ESTIMATION OF PLANETARY WAVE PARAMETERS FROM THE DATA OF THE 1981 OCEAN ACOUSTIC TOMOGRAPHY EXPERIMENT

by

Ching-Sang Chiu
B.S., 1979 Northeastern University

SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

and the

WOODS HOLE OCEANOGRAPHIC INSTITUTION

August 1985

Signature of Author.....
Department of Ocean Engineering, Massachusetts Institute of Technology and the Joint Program in Oceanography/Oceanographic Engineering, Massachusetts Institute of Technology/Woods Hole Oceanographic Institution, August 1985

Certified by .....
Yves J.F. Desaubies, Thesis Supervisor

Accepted by .....
William D. Grant, Chairman, Joint Committe for Oceanographic Engineering, Massachusetts Institute of Technology/Woods Hole Oceanographic Institution

ESTIMATION OF PLANETARY WAVE PARAMETERS FROM THE DATA OF THE 1981 OCEAN ACOUSTIC TOMOGRAPHY EXPERIMENT

by

Ching-Sang Chiu

Submitted to the Massachusetts Institute of Technology/Woods Hole Oceanographic Institution Joint Program in Oceanographic Engineering in August 1985 in partial fulfillment of the requirements for the Degree of Doctor of Science

ABSTRACT

Using the maximum-likelihood estimation method and minimization techniques, quasi-geostrophic wave solutions were fitted to the observations of the 1981 Ocean Acoustic Tomography Experiment. The experiment occupied a 300 km square area centered at 26°N, 70°W, and had a duration of ~80 days. The data set consisted of acoustic travel-time records, temperature records and CTD profiles, obtained from the acoustic tomographic array, moored temperature sensors and recorders, and ship surveys, respectively. While the latter two were conventional spot measurements, the former corresponds to integral measurements of the temperature (or sound-speed) field.

The optimal fit to the data corresponded to 3 waves in the first baroclinic mode, evolving under the presence of a westward mean flow with vertical shear. The flow was estimated to be weak (~2 cm/s), but it changed the wave periods significantly by producing large Doppler shifts. The waves were dynamically stable to the mean flow, had weak nonlinear interactions with each other and did not form a resonant traid; thus they constituted a fully linear solution.

Evidence for the existence of the waves was strongly supported by the high correlation (~0.9) between the data and the fit, the large amount of signal energy resolved (~80 percent), the excellent quality of the wave-parameter estimate (only about 10 percent in error), and the general agreement between the observations and quasi-geostrophic linear dynamics.

Thesis Supervisor: Dr. Yves J.F. Desaubies
Associate Scientist, Woods Hole Oceanographic Institution, Woods Hole, MA.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I would like to thank my thesis advisor, Yves Desaubies, for his excellent teaching, for his constant support and encouragement, and for his investment of time and energy. My sincere appreciation for his valuable comments on the preliminary drafts of this thesis. I also wish to show my graditude to the other members of my thesis committee for their discussions and help on various aspects. In particular, Carl Wunsch read the preliminary drafts and gave important suggestions and comments; Bob Spindel taught me acoustic tomography; Arthur Baggeroer offered good academic advice at M.I.T.. My financial support for the first two years came from the Education Office at W.H.O.I.. My dissertation research was supported by ONR Grant N00014-82-C0019.

I am grateful to Jim Lynch. He served as chairman in my thesis defense committee, and donated much of his spare time to carefully proof the final version of this thesis. Special thanks to Maxine Jones and Stanley Rosenblad for the help they gave whenever I had difficulty with the computer. My appreciation to Ann Henry for typing the review papers for my part-two general examination.

I am indebted to the Ocean Tomography Group, and especially to Dave Behringer, Bruce Cornuelle, Bob Spindel and Carl Wunsch. They generously handed me the data set, and trusted me that I would do something with it. I hope that they are not disappointed.

Finally, I want to thank my family and Amy's family for the continual support and encouragement. I am especially grateful to Amy for all the sacrifices she has made during these years.

CHAPTER 1

INTRODUCTION

Over the last two decades, several vigorous research programs have been conducted by scientists to study oceanic mesoscale variability. As a consequence, a more detailed and realistic description of the ocean circulation has been obtained. Much of the knowledge of the variability has been obtained from extensive experiments such as POLYGON, 1970 (Brekhovskikh et al., 1971), MODE-0, 1971-1972, MODE-1, 1973 (MODE Group, 1978) and the recent POLYMODE, 1974-1978 (U.S. POLYMODE Organizing Committee, 1976) in which multi-moorings and a variety of instruments were deployed to observe the four-dimensional fields of current and density at mid-latitudes in the North Atlantic. Today, it is well-known that mesocale fluctuations that are often called 'eddies' are energetically dominant and exist everywhere in open oceans. Even close to land, numerous observations of trapped mesoscale motions have also been reported (Longuet-Higgins, 1968, Wunsch, 1972, and Hogg, 1980).

Besides being the most dominant feature in the ocean, eddies interact with the mean circulation through the processes of energy cascades to larger-scale flows (Rhines, 1975) and barotropic and baroclinic instabilities (Pedlosky, 1979), and they transport heat and salt effectively by their intense flow field. Therefore, the knowledge of eddy dynamics is of fundamental interest to physical

oceanographers in understanding the general circulation. Furthermore, the research is also of great significance to meteorologists and marine scientists in other disciplines, since ocean eddies can influence the long-term climate on earth through air-sea interaction, transport chemicals, and relocate biological matter.

Mesoscale eddies are characterized by periods of 50 to 100 days, horizontal scales of order 100 km and vertical scales comparable to the depth of the ocean (Richman et al., 1977, and McWilliams, 1979). In places where the flow field is strong, for example in regions close to the major currents, the fluctuations are nonlinear turbulent motions. However, it is conceivable that the fluctuations can be wave-like and dispersive in places that are relative calm, because the linearized equation of mesoscale motion, that is the linearized quasi-geostrophic potential vorticity equation, does admit planetary wave solutions (LeBlond and Mysak, 1978, and Pedlosky, 1979). Furthermore, the wave solution does exhibit behavior that is consistent with some observations, for example, westward phase propagation.

Literature on the theory of planetary waves is abundant, but only slight observational evidence for their existence in open oceans exists. Perhaps, the most striking evidence to date was found by McWilliams and Robinson (1974), and McWilliams and Flierl (1976), by fitting waves to the POLYGON observations and the MODE-array data, respectively. POLYGON was conducted by the USSR in

the tropical North Atlantic during the spring and summer of 1970. The array, which centered at 16°30'N, 33°30'W, measured the eddy currents for several months from moored current meters and hydrography. The data was analyzed and presented by Koshlyakov and Grachev (1973). They inferred that a single, anti-cyclonic eddy, a few hundred kilometers in diameter, traversed the array during the experiment, and synthesized their observations in terms of a moving elliptical cylinder representing the locus of maximum horizontal current at each depth. McWilliams and Robinson (1974) fitted planetary waves in a two-layer model to the descriptive synthesis, in which the free parameters, that is the wave amplitudes and wavenumbers were determined from the major and minor axes, the orientation angle and the maximum orbital speed of the ellipse. It was found that the synoptic structure and propagation of the ellipse were well matched by a pair of baroclinic waves with equal pressure amplitudes. However, the POLYGON wave fit was highly subjective and might not be optimal due to the fact that the number of waves was arbitrarily chosen and the observations used were not the actual data themselves. The lack of actual data has prevented McWilliams and Robinson from making a quantitative assessment of the wave model.

The Mid-Ocean Dynamics Experiments MODE-0 and MODE-1 were conducted jointly by the USA and UK in an approximately 400 km square region centered at 28°N, 69°40'W, again in the tropical North Atlantic. MODE-0 was a collection of several pilot studies that were carried out between 1971 and 1972 to identify the energy level,

and space and time scales of the mesoscale motion. It was then followed by MODE-1, which was a more comprehensive experiment designed to provide an accurate four-dimensional mapping of a mid-ocean eddy during the spring of 1973. Several combinations of barotropic and baroclinic waves in a continuous ocean model were fitted to the MODE-0 and MODE-1 data sets by McWilliams and Flierl (1976). While the MODE-0 data set contained only current-meter records having durations of from 1 to 3 months, the MODE-1 data set was much larger and more uniform in space and time, having a duration of 4 months. It also contained different types of observations, i.e. from current meters, moored temperature sensors, hydrographic stations and float tracks. In the fitting process, the free wave parameters were chosen optimally to minimize a quadratic error norm for the differences between the data and fit. While the best MODE-1 fit consisted of a pair of waves in the barotropic mode and a pair of waves in the first baroclinic mode, the best MODE-0 fit consisted of a pair of barotropic waves only. Both MODE wave fits were fairly successful, having correlations of ~0.7 with the data and accounting for ~1/2 of the observed signal energies, i.e. ~70 percent of the signals (rms). However, the MODE-1 fit corresponded to an inconsistent linear solution: nonlinear wave-wave interactions within the fit were predicted by the weakly nonlinear theory to be strong but were not found in the data. Thus, there remains some doubt as to whether planetary waves truely existed during MODE-1, and more fundamentally perhaps, whether planetary

wave propagation is a typical dynamical phenomenon in that part of the ocean.

The purpose of this dissertation is threefold. First, it reinvestigates the existence of planetary waves in the tropical North Atlantic. This time, the investigation is done by trying to detect the wave signals from the acoustic and spot observations made in the 1981 Ocean Acoustic Tomography Experiment (The Ocean Tomography Group, 1982), and in doing so, the wave dynamics in the region which is centered at 26°N, 70°W (which will be referred to as the tomographic region) is also investigated. Second, it examines the performance of the acoustic-tomographic observational system, the spot-observational system and the combination of the two systems, as deployed in the experiment, in observing the waves and also in mapping the ocean. Third, it explores the possibility of using acoustic tomography to provide adequate large-scale monitoring in the absence of the tracking of the motion of the acoustic moorings.

The investigation of the existence and dynamics of planetary waves involves analyzing the fits of different but plausible wave-propagation models to different types of observations of sound-speed or temperature perturbations, made by the CTD casts, temperature sensors, temperature-pressure recorders and the acoustic tomographic array deployed in the experiment. The hope is to be able to detect the waves and, at the same time, determine the correct wave dynamics in the fitting process by comparing the

quality of the different wave-model fits. Due to the insufficiency of explicit current measurements which came from only two horizontal locations, some deficiences will persist in our investigation. For example, we cannot observe the barotropic waves and explore the thermal wind relation between the wave-induced current and density perturbations.

The technique of fitting used here is procedurally similar to that used by McWilliams and Flierl (1976), corresponding to the minimization of a quadratic error norm between the data and the wave fit, that is a weighted sum of products of residuals. However, a fundamental difference is that, while they have defined their error norm by choosing the weighting factors in a subjective manner as to give equal weighting to each subset of data of the same type, we have constructed our norm by adopting the idea of maximum likelihood from the stochastic framework, i.e. the weighting factors are the reciprocals of the noise variances. The appeal of using statistical approaches is that the meaning of a wave fit being the 'optimal' or 'best' can be explicitly defined in terms of statistical conditions. Another difference is that our fitting involves acoustic observations that correspond to integral measurements of the field in addition to spot observations.

We must give credit to The Ocean Tomography Group who provided the data. The experiment was conducted by them primarily for the testing of 'Ocean Acoustic Tomography' which is a pure acoustic inverse scheme for monitoring large-scale fluctuations in ocean

basins. The innovative idea of ocean tomography was first introduced by Munk and Wunsch (1979) and the scheme is analogous to the medical tomographic procedure CAT scan. A typical mid-ocean tomographic system, as described by Munk and Wunsch and deployed in the experiment, consists of a sparse horizontal array of moored mid-water acoustic sources and receivers that surrounds a large area of the ocean under study, so that by exploiting the properties of sound propagation in the SOFAR channel, such as low attenuation and multipath arrivals, the entire volume can be remotely sensed, horizontally, vertically, and temporally with large-scale resolution by using repeated acoustic transmissions. Thus, through mathematical modeling of the relation between oceanic and acoustical fluctuations, the four-dimensional sound-speed perturbation field should be reconstructable based on the observed perturbations of the multipath arrival times using inverse techniques. Superior to traditional spot-measurement techniques, acoustic tomography can monitor a larger region and provide a larger database with fewer moorings, and its averaging (integrating) process can filter out undesirable small-scale oceanic features automatically. Furthermore, unlike shipboard surveys, it can map the ocean instantly and the mapping can be done frequently. These advantages of cost effectiveness and high temporal resolution are some of the appeal of acoustic tomography. However, the acoustical scheme depends critically on the stability, identification and resolution of multipaths over long distances. These have been verified by

Spiesberger et al. (1980) and Spindel and Spiesberger (1981) in preliminary experiments.

The 1981 experiment was the first field test of a full tomographic system for mapping the ocean at mesoscale resolution. In order to evaluate the performance of the system, the experimental region was also measured with traditional techniques by The Ocean Tomography Group during the same time. The idea was to provide a basis for comparison. The tomographic system in a linear form was later 'inverted' for the three-dimensional sound-speed perturbation fields, independently of time and only with acoustic data, by Cornuelle (1983) and Cornuelle et al. (1985). Because the daily tomographic maps do compare favorably with the ship-based objective maps, they have demonstrated the practicality of acoustic tomography for mesoscale monitoring. Here, our principal objective is to investigate the existence and dynamics of planetary waves; therefore, in order to obtain the best estimate of the wave parameters and wave dynamics, we have incorporated the spot measurements of temperature as well as the integral measurements (that is the acoustic travel-time data) in our estimate.

The inversions of the data performed in this study are for the retrieval of the planetary wave parameters and the planetary wave field, and are intrinsically different from those previously done by Cornuelle (1983) and Cornuelle et al. (1985). The originality of our inversions lies in that they give a time-dependent estimate of the unknown field, the system involved is nonlinear with respect to

the unknown parameters, and contains both acoustic and traditional (spot) observations. Specificically, the system is 'inverted' for the four-dimensional sound-speed perturbation field subject to the different dynamical constraints constituted by the plausible models of wave propagation. The inversions, therefore, besides producing maps of the ocean structure, also test different wave dynamics against the data for consistency and optimality. Due to the nonlinear nature of our inverse problem, standard linear techniques such as Singular Value Decompositions and Gaussian Eliminations are not applicable, so that we use iterative descent minimization techniques to solve the problem.

In order to observe the waves, the forward problem of how the observations of the dynamical field are related to the evolution of the waves under different dynamical conditions must first be resolved. This subject is pursued in Chs. 2 and 3. In Ch. 2, we examine the theory of planetary waves by reviewing the literature. We review the evolution of the waves at mid-latitude, and under the possible effects of weak mean current, small bottom slope and weakly nonlinear wave-wave interaction. An objective is to illustrate that the space and time behavior is constrained by the modal dispersion relationship and characterized by the wave parameters: wavenumbers, wave amplitudes, modal amplitudes of the mean flow and growth rates. In Ch. 3, we develop the model equations that relate the data to the wave parameters that characterize the wave and mean-flow induced sound-speed perturbation. We also describe the filtering

and reduction of the data prior to the inversions. Furthermore, we present three plausible dynamical models of the induced sound-speed perturbation, which have been fitted to the data to estimate the wave dynamics.

In Ch. 4, we discuss the general parameter-estimation or inverse problem. The goals are to relate and unify some commonly used estimation methods, deterministic or stochastic, and to show that there is a general estimation procedure, common to all the methods considered, to obtain the optimal solution. The procedure corresponds to the minimization of an objective function of a weighted sum of products of residuals, that is a quadratic error norm. We also discuss the error variance of an estimate and some widely used numerical techniques for minimization. We further present some simple measures of goodness of the fit for appraising models.

Using a gradient method for minimization (Fletcher and Powell, 1963), the wave parameters of each of the three plausible wave models were estimated. This corresponds to wave fitting, and in order to estimate the number of waves, a range of one to five waves was assumed for each model in the fitting. The results of the wave fits and the identification of the optimal model and number of waves are described and discussed in Ch. 5. Furthermore, the error variance of the estimated wave and mean-flow induced sound-speed perturbation, associated with the error of the optimal estimate of wave parameters, is analysed.

In Ch. 6, we first summarize the results of the wave fits and comment on the dynamics, linearity and stability of the waves observed. We then compare this wave fit with the MODE wave fits, and from the results of the three wave fits, we make general statements on the wave dynamics in the area occupied by the experiments. We also compare the tomographic inverse method of Cornuelle (1983) and Cornuelle et al. (1985) with our method, and analyse the ability of the acoustic, spot and mixed observational systems in observing the waves and mapping the ocean. We then make concluding remarks on the investigation.

The motion of the acoustic moorings, if not tracked, can be misinterpreted as oceanic fluctuations in a tomographic inversion. However, for economical reasons, it is highly desirable to know whether reliable acoustic mapping of the ocean structure can still be generated without the deployment of navigational systems for tracking mooring motions, but rather through parameterization of the mooring motions, as was done by Cornuelle (1983). As a secondary contribution by this dissertation, a study of this engineering problem is presented in Ch. 7.

In Ch. 7, we derive bounds on the error of the tomographic sound-speed estimate in the presence of untracked mooring motions. An important result shows that the error variance of the estimate is practically invariant with the size of mooring motion but is almost always reaching the upper variance bound. The implication is that, given a priori information about the field, the geometry of the

tomographic array, and the noise level, the upper bound can be evaluated to give an indication about whether it will be necessary to track the moorings before the deployment.

Not to bore the readers who are experts on the subjects of planetary waves and parameter estimation, or only interested in the data-model relations and the estimation results, we take this opportunity to inform them to skip Chs. 2 and 4 in their reading. These two chapters contain only review material. The literature on the two subjects is vast, and our only excuse for writing Chs. 2 and 4 is to define the mathematical notation used in this thesis. New material and results are contained in Chs. 5, 6 and 7, and in part of Ch. 3. The acoustic forward problem considered in Ch. 3 has previously been studied by Munk and Wunsch (1979), Cornuelle (1983) and many others, and the reason for the redundancy here is just to make this presentation of the forward problem a complete one. New material in Ch. 3 are the results of the analytical-mode decompositions of the CTD data, the use of the modal decompositions as a data reduction scheme and a demonstration of the transparency of the higher modes to acoustic measurements.

CHAPTER 2

MESOSCALE PERTURBATIONS AND WAVE MOTIONS

In the open ocean, the largest portion of the total kinetic energy is contained in the mesoscale frequency band. Mesoscale perturbations or eddies have characteristic flow speeds of centimeters per second, horizontal length scales of hundreds of kilometers, vertical length scales comparable to the depth of the ocean, typical oscillation periods of months, and westward phase velocities. Over nonsteep and smooth bottom topography, eddy currents are basically horizontal, the momentum balance is almost geostrophic and the local dynamics are governed by the law of conservation of quasigeostrophic potential vorticity.

Away from intense mean currents, lateral boundaries and steep bottom topography, dispersive planetary (or Rossby) waves of low frequencies and large length scales can propagate due to the latitudinal variation of the coriolis parameter. These waves are solutions of the linearized equation of the conservation of quasigeostrophic potential vorticity. The linearization is valid when the ratio of the wave period to the advective time is small compared to unity. Under such circumstances, mesoscale fluctuations in the flow field and the density field are direct consequences of the propagation of planetary waves; the density fluctuations are in turn related to temperature and sound-speed perturbations.

This chapter is intended to examine, by reviewing the literature, the dynamics of planetary waves, and the underlying dynamical and geometrical approximations used on the basic equations of motions. Sources of reference are LeBlond and Mysak (1978) and Pedlosky (1979) for the scaling analysis on the basic equations, the derivation of the quasigeostrophic potential vorticity equation and the general theory of planetary waves, Flierl (1978) for the orthonormalization of the quasigeostrophic potential vorticity equation and the derivation of the horizontal-structure equations associated with the normal modes, and Longuet-Higgins et al. (1967) for the theory of resonant wave-wave interactions. The mechanisms for wave generation and dissipation will not be considered, the focus will be on the evolution of planetary waves at mid-latitude, under the influnence of the earth's rotation, and under the effects of weak mean currents, small bottom slopes and weakly nonlinear wave-wave interations. The goals are to derive relations between perturbed dynamical variables and wave-parameters such as wave-amplitudes, wavenumbers and wavefrequencies, and most important of all, to carefully study how planetary waves propagate and interact. Our knowledge of mesoscale variability can be increased if some dynamical variables are measured or remotely sensed and wave parameters are then estimated.

2.1 Governing Equations For Mesoscale Motions

2.1.1 Basic Laws Of Conservation

The conservation laws for an unforced, incompressible, nondiffusive (in both heat and salt) and invicid ocean model are (LeBlond and Mysak, 1978)

$$\frac{d\underline{v}}{dt} - 2\underline{\omega}\times\underline{v} = -\frac{1}{\rho}\nabla p + \underline{g}, \tag{2.1}$$

$$\frac{d\rho}{dt} = 0, \tag{2.2}$$

and

$$\nabla \cdot \underline{v} = 0, \tag{2.3}$$

where d/dt is the total derivative, all the dynamical variables are functions of time and space, $\underline{v}$ is the velocity vector of fluid particles relative to the rotating frame associated with the earth that has a constant angular velocity vector $\underline{\omega}$ (its magnitude is $\omega \sim 7.3\times10^{-5}$ rad/s), ρ and p are the density of the fluid and the pressure acting on it, respectively, and the $\underline{g}$ vector is the acceleration of gravity (its magnitude is $g \sim 9.81$ m/s^2). The

conservation of momentum is expressed in (2.1), (2.2) is a statement regarding the thermodynamic properties of nondiffusivity and incompressibility, (2.3) expresses continuity (conservation of volume) and is a combined result of conservation of mass and (2.2).

In the static state where $\underline{v}=0$ and $\rho=\rho_0$ is a function of depth -z or the radial coordinate only, the hydrostatic pressure p_0 is related to ρ_0 by

$$\frac{dp_0(z)}{dz} = - \rho_0(z)\, g. \tag{2.4}$$

We would like to point out that the static state is generally different from the mean state, i.e. they would be the same only when there is no mean motion. In a nonstatic state where the fluid has motion, the pressure and density depart from hydrostatics to become $p=p_0+p'$ and $\rho=\rho_0+\rho'$, and (2.1) and (2.2) can be rewritten as

$$\frac{d\underline{v}}{dt} - 2\underline{\omega}\times\underline{v} = - \frac{1}{\rho^*} \nabla p' + \frac{\rho'}{\rho^*} \underline{g} \tag{2.5}$$

and

$$\frac{d\rho'}{dt} + w \frac{d\rho_0}{dz} = 0, \tag{2.6}$$

respectively, where w is the vertical or radial velocity. In (2.5), ρ is replaced by a constant reference density ρ^*~1 g/ml (the Boussinesq approximation) because the variation of ρ in both time and space is only about one percent througout the ocean, hence the replacement would insignificantly alter the coriolis and inertial forces.

2.1.2 Scalings And Approximations

Scaling analysis can be employed to simplify the complicated basic set of equations (2.3), (2.5) and (2.6) to a set that describes only mesoscale motions at mid-latitude. The method of simplification which is described in detail by Pedlosky (1979) consists in, as a first step, the transformation from the spherical coordinate system to one with x,y and z coordinates representing the eastward, northward and upward distances, repectively, measured from the transformed origin located on the surface of the ocean, at a latitude θ_0 where the area under study is centered. The transformation includes the Taylor expansions of the trigonometric functions of latitude θ, which appear in the equations because of sphericity, about θ_0 in powers of x and y. The components of $\underline{v}$ are now u, v and w corresponding to the x, y and z directions, respectively. As a second step, the independent variables are scaled and the dependent (dynamical) variables are normalized so that a set of nondimensional equations is obtained. The scalings

and normalizations are done by using observed characteristic lengths, times and flow speeds, and also by using observed or estimated magnitudes of w, p' and ρ'. The quantities used for the scalings and normalizations are shown in the second row of table (2.1). At mid-latitude, a typical horizontal length scale is L~100 km, a typical vertical length scale is H~1 km and a characteristic horizontal flow speed is U~5 cm/s. From continuity, an estimate of an upper bound for the magnitude of w is UH/L and this quantity is used for its scale. It is important to point out the way that p' and ρ' are scaled is due mainly to our perception that the motions are almost geostrophic and hydrostatic.

Next, the scaled dynamical variables are expanded as perturbation series in powers of a small parameter ε. Then equations that describe the temporal and spatial behavior of the nonvanishing leading terms in the expansions are sought. The small parameter is the Rossby number and, approximately, two other important small geometrical ratios:

$$\varepsilon = U/f_0 L \sim L/R \sim H/L \sim 10^{-2}, \tag{2.7}$$

where $R \sim 6.36 \times 10^3$ km is the earth's radius and $f_0 \sim 10^{-4}$ rad/s is the coriolis parameter $f = 2\omega \sin\theta$ evaluated at θ_0. The smallness of the Rossby number $U/f_0 L$ and the aspect ratio H/L indicates that the flow is predominantly geostrophic and horizontal. The neglect of higher-order terms emphasizes that our interest is in local

dynamics, with the localization in space explicitly indicated by the ratio L/R.

table 2.1

Summary Of Orders Of Magnitudes And Scales

variables	x,y	z	t	u,v	w	p'	ρ'
scaling or normalizing factor	L	H	L/U	U	UH/L	$\rho * f_0 UL$	$\rho * f_0 UL/gH$
order of magnitude				U	$\varepsilon UH/L$	$\rho * f_0 UL$	$\rho * f_0 UL/gH$
order of magnitude of error in quasigeostrophic solution				εU	$\varepsilon^2 UH/L$	$\varepsilon\rho * f_0 UL$	$\varepsilon\rho * f_0 UL/gH$

2.1.3 Quasigeostrophy

After collecting nondimensional terms in the equations with like powers of ε, we find that to the lowest order in ε (that is order ε^0), the motion is geostrophic (equations will be put back in dimensional forms),

$$(u,v) = \frac{1}{\rho^* f_0} \left(-\frac{\partial p'}{\partial y}, \frac{\partial p'}{\partial x} \right), \tag{2.8}$$

hydrostatic,

$$\frac{\partial p'}{\partial z} = -\rho' g, \tag{2.9}$$

horizontally nondivergent, and the zeroth-order w vanishes. Note that $p'/\rho^* f_0$ is a geostrophic (zeroth-order) stream function and $\nabla_H^2 p'/\rho^* f_0$ is the geostrophic (zeroth-order) relative vorticity as indicated by (2.8); $\nabla_H^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2$ is the horizontal Laplacian. Equations (2.8) and (2.9), in a sense, are not too interesting because they do not provide any new information nor information regarding the evolution of the perturbations in time. However, it is clear that w is more accurately of order $\varepsilon UH/L$, which is even smaller than the original estimate. The precise order of magnitudes of the dependent variables are summarized in the third row of table (2.1).

Although w is very small (a first-order quantity), it must be taken into account in order to study mesoscale dynamics. In fact, by considering also the first-order equations in ε, it is found that changes in the vertical component of the zeroth-order absolute vorticity (planetary plus relative vorticities) along a particle's path line are produced solely by the streching of vortex tubes or the small divergence of the horizontal flow $\partial w/\partial z$:

$$\frac{d_H}{dt}\left(\frac{1}{\rho^* f_0}\nabla_H^2 p' + f\right) = f_0 \frac{\partial w}{\partial z}, \qquad (2.10a)$$

where

$$\frac{d_H}{dt} = \frac{\partial}{\partial t} + u\frac{\partial}{\partial x} + v\frac{\partial}{\partial y} = \frac{\partial}{\partial t} + \frac{1}{\rho^* f_0}\left(-\frac{\partial p'}{\partial y}\frac{\partial}{\partial x} + \frac{\partial p'}{\partial x}\frac{\partial}{\partial y}\right). \qquad (2.10b)$$

As a result of the geometrical scalings and the neglect of higher-order terms, the vertical planetary vorticity or the coriolis parmeter f in (2.10a) is evaluated locally as

$$f = f_0 + \beta y, \qquad (2.11)$$

where $\beta = 2\omega\cos\theta_0/R$ ($\sim 2\times10^{-8}$ rad/s/km) is the latitudinal gradient of f evaluated at θ_0. It is also obtained that w is related to p' by

$$w = \frac{d_H}{dt}\left(\frac{-1}{\rho^* N(z)^2}\frac{\partial p'}{\partial z}\right) \qquad (2.12)$$

where

$$N(z)^2 = -\frac{g}{\rho^*}\frac{\partial \rho_0(z)}{\partial z} \quad ; \tag{2.13}$$

N(z) is the Brunt-Vaisala frequency that characterizes the stability of the water column and is assumed to be known from density measurements. Obviously, the vertical displacement of isopycnal surfaces (or isothermal surfaces or surfaces of constant sound speed) is, from (2.12),

$$\eta = \frac{1}{\rho^* N^2}\frac{\partial p'}{\partial z}. \tag{2.14}$$

We would like to add that in collecting terms to like orders, we have used the fact that the Burger's number $(HN/Lf_0)^2$ is of order one since $N^2 \sim 10^{-5}$ $(rad/s)^2$.

The consideration of quasigeostrophy, that is the small deviation from geostrophy or the small w, leads further to the derivation of a single equation for the stream function in a closed form (the equation is obtained by combining (2.10) and (2.12)):

$$\frac{d_H}{dt}\left[\left(\nabla_H^2 + \frac{\partial}{\partial z}\frac{f_0^2}{N^2}\frac{\partial}{\partial z}\right)p'\right] + \beta\frac{dp'}{dx} = 0. \tag{2.15}$$

Since it is known that the potential vorticity $(\nabla \times \underline{v} + 2\underline{\omega}) \cdot \nabla \rho$ is conserved following a fluid particle in an incompressible, adiabatic (invicid and nondiffusive) and unforced ocean (the proof can be found in Leblond and Mysak, 1978), it is interesting to point out that (2.15) is simply a statement of this conservation law but following from the applications of the geostrophic, hydrostatic, geometrical and Boussinesq approximations. Therefore, the governing equation for mesoscale motions is the conservation of quasigeostrophic potential vorticity.

We have already derived relations between p' and other dynamical variables. Once (2.15) is solved for p' with the appropriate boundary conditions, other dynamical variables are then known from (2.8), (2.9) and (2.12). The solutions are not exact but are zeroth-order approximations for p',u,v and ρ', and a first-order approximation for w, hence they are accurate to within about 100ε percent, that is about one percent. The order of magnitudes of the errors in the quasigeostrophic solutions are summarized in the fourth row of table (2.1).

2.2 Boundary Conditions

The boundary conditions are the continuity of pressure and displacement across the disturbed ocean surface at $z=s(x,y,t)$, and the vanishing of the normal velocity at the rigid bottom at $z=-D+b(x,y)$; D is the nominal depth of the ocean and $D>>|s|$ and $|b|$. However, it is desirable to scale and approximate these conditions so that they can be replaced by a simplified but consistent version that applies to p' at $z=0$ and $z=-D$ instead. Otherwise, it would be a very difficult task to solve (2.15). The simplifications will be detailed in the following sections.

2.2.1 The Surface

The exact conditions are, at $z=s$,

$$p_0(z) + p'(x,y,z,t) = p_a, \tag{2.16a}$$

and

$$w = ds/dt \ . \tag{2.16b}$$

The atmospheric pressure p_a can be assumed constant as far as the ocean is concerned, because the magnitude of the variation of p_a is much smaller and the length scale of variation is much larger.

After substituting the Taylor expansions of the dynamical variables about z=0 in powers of s in (2.16), and then dropping nonlinear terms in s, p' and w (so that only the largest terms are kept), we obtain, at z=0,

$$p' \sim \rho_0 g s \tag{2.17a}$$

and

$$w \sim d_H s/dt \tag{2.17b}$$

with the uses of (2.4) and the identity $p_0=p_a$. The above two equations can be combined to give, at z=0,

$$w \sim \frac{d_H}{dt}\left(\frac{p'}{\rho_0 g}\right). \tag{2.18}$$

An order of magnitude analysis (by using table (2.1)) shows that the R.H.S. of (2.18) is of order $\varepsilon(L^2f_0^2/gH)(UH/L)$, but it is also of order $\varepsilon^2(UH/L)$ since $L^2f_0^2/gH$ (estimated with the typical values of L, f_0, g and H) is approximately equal to ε. In conclusion, the R.H.S. of (2.18) that introduces only a second-order correction to w can be consistently discarded without affecting the quasigeostrophic solution. The result is the rigid-lid approximation, that is

$$w(x,y,0,t)=0, \tag{2.19}$$

or equivalently,

$$\frac{d_H}{dt}\left(\frac{-1}{\rho^* N^2}\frac{\partial p'}{\partial z}\right) = 0 \qquad \text{at } z=0, \tag{2.20}$$

as obtained by using (2.12).

2.2.2 The Bottom

The exact boundary condition at the bottom can be written as

$$w = u\,\partial b/\partial x + v\,\partial b/\partial y \qquad \text{at } z=-D+b. \tag{2.21}$$

Substitution of the linear expansions of u, v and w about z=-D in b and dropping the nonlinear terms in w and b in (2.21) gives

$$w \sim u\,\partial b/\partial x + v\,\partial b/\partial y \qquad \text{at } z=-D. \tag{2.22}$$

In order for quasigeostrophic theory, which requires w to be of order $\varepsilon UH/L$, to remain valid, we must restrict the magnitudes of the slopes to be approximately equal to or smaller than $\varepsilon H/L$. On the

other hand, if the magnitudes of the slopes approach $\varepsilon^2 H/L$, we can consistently set w=0 at z=-D without affecting the solution.

In using (2.8), (2.12) and (2.22), the condition for p' can be written as

$$\frac{d_H}{dt}\left(\frac{-1}{\rho^* N^2}\frac{\partial p'}{\partial z}\right) = \frac{1}{a^* f_0} J(p',b) \qquad \text{at } z=-D, \tag{2.23a}$$

where

$$J(p',b) = \frac{\partial p'}{\partial x}\frac{\partial b}{\partial y} - \frac{\partial p'}{\partial y}\frac{\partial b}{\partial x} \tag{2.23b}$$

is the Jacobian operation.

2.3 Normal Modes

Ultimately, we want to solve the nonlinear quasigeostrophic potential vorticity equation (2.15) subject to the nonlinear boundary contitions in (2.21) and (2.23). However, if the method of separation of variables is used to solve the linearized problem in the case of a flat bottom, a set of z-dependent eigenfunctions $f_i(z)$, called the normal modes for p', are found. They obey the vertical (structure) equation:

$$\frac{d}{dz}\left(\frac{f_0^2}{N^2}\frac{df_i}{dz}\right) + \lambda_i f_i = 0;\ i=0,1,2,\ldots, \tag{2.24a}$$

with

$$\frac{df_i(0)}{dz} = \frac{df_i(-D)}{dz} = 0;\ i=0,1,2,\ldots, \tag{2.24b}$$

where λ_i is the corresponding eigenvalue. $\lambda_i^{-1/2}$ is called the radius of deformation of the ith mode. Since the $f_i(z)$'s constitute a complete set of orthogonal functions of z, the solution for the nonlinear problem can be cast as

$$p' = \sum_i p_i'(x,y,t) f_i(z). \tag{2.25}$$

In view of (2.8), (2.9) and (2.14), we can also write

$$(u,v) = \left[\sum_i u_i(x,y,t) f_i(z) , \quad \sum_i v_i(x,y,t) f_i(z) \right], \tag{2.26}$$

$$\rho' = \sum_i D\, \rho'_i(x,y,t) f'_i(z) \tag{2.27}$$

and

$$\eta = \sum_i \eta_i(x,y,t) h_i(z), \tag{2.28}$$

where $f_i' = df_i/dz$ and $h_i = Df_0^2 f_i'/N^2$. Futhermore, the modal-amplitude functions are related by

$$(u_i, v_i) = \frac{1}{\rho^* f_0} \left(-\frac{\partial p_i'}{\partial y}, \frac{\partial p_i'}{\partial x} \right), \tag{2.29}$$

$$\rho'_i = -p'_i/gD \tag{2.30}$$

and

$$\eta_i = -p'_i/\rho^* f_0^2 D. \tag{2.31}$$

Because the vertical displacement η is intimately related to the

commonly observed sound speed (or temperature), it is used here in place of w.

In (2.25) to (2.28), the vertical structure of p', (u,v), ρ' and η is decomposed into normal modes. The modal decompositions can be achieved by first solving the Sturm-Liouville problem in (2.24) for the $f_i(z)$'s (the normal modes for p') and λ_i's with a known N^2, one then evaluates the $f'_i(z)$'s and $h_i(z)$'s with $f_i(z)$'s, accordingly. On the other hand, one can first obtain the normal modes for η by solving an equivalent eigenvalue problem:

$$\frac{d^2}{dz^2} h_i + \frac{\lambda_i}{f_0^2} N^2 h_i = 0; \quad i=0,1,2,\ldots, \tag{2.32a}$$

with

$$h_i(0) = h_i(-D) = 0; \quad i=0,1,2,\ldots. \tag{2.32b}$$

This is done by Mooers (1975) in his investigation of linear waves and the corresponding sound-speed perturbations in a flat-bottomed ocean with no mean flow. Equations (2.32) can be derived directly from (2.24).

The sound-speed perturbation field $\delta c(x,y,z,t)$ is created by the vertical displacement of the surfaces of constant sound speed:

$$\delta c = -\eta \left[\frac{d}{dz} c_0(z) - \frac{d}{dz} c_A(z) \right], \tag{2.33}$$

where c_0 and dc_A/dz are the mean profiles of sound speed and its adiabatic gradient arising from the adiabatic expansion or compression of a rising or sinking volume of fluid, respectively. The quantity in the bracket is the potential gradient of sound speed (Flatte et al., 1979). Unlike the case for ρ', compressibility must be taken into account in the evaluation of δc because the adiabatic gradient of sound speed is not small in comparison with its potential gradient and adiabatic gradients do not contribute to fluctuations. A modal representation of δc is

$$\delta c = - \sum_i \eta_i(x,y,t)\, f_0 g_i(z), \tag{2.34a}$$

with

$$f_0 g_i(z) = h_i(z) \frac{d}{dz}[c_0(z) - c_A(z)]. \tag{2.34b}$$

$f_0 g_i$ can be interpreted as the vertical anomaly of sound speed per unit displacement of the ith mode. The buoyancy frequency profile N(z) measured during the tomographic experiment in 1981 is plotted in Fig. 3.3, from which the first three baroclinic normal modes for p' (or (u,v)) are evaluated and plotted in Fig. 3.4. The corresponding normal modes for η and δc are also evaluated, renormalized to have maxima of unity and plotted in Fig. 3.5 and 3.6, respectively.

The description of the modal solution for quasigeostrophic motions would be incomplete without the horizontal (structure) equations that govern the modal-amplitude functions $p_i(x,y,t)$. Briefly, (2.15) is multiplied by $f_n(z)$ and p' is replaced by its modal representation in (2.25). Integration along z is then performed to eliminate the z-dependence of the equation. This elimination is accomplished with the use of the orthonormality condition

$$\frac{1}{D}\int_{-D}^{0} f_i(z)f_n(z)dz = \delta_{in}, \tag{2.35}$$

where δ_{in} is the krononeker delta. For more details regarding the procedure for the orthonormalization, one can consult Flierl (1978). The resulting equations are

$$\left[\frac{\partial}{\partial t}(\nabla_H^2-\lambda_n)+\beta\frac{\partial}{\partial x}\right]p_n' + \frac{1}{\rho^* f_0}\sum_i\sum_j \varepsilon_{ijn}J[p_i', (\nabla_H^2-\lambda_n)p_j'] = \frac{-f_0}{D}f_n(-D)\sum_i J(p_i',b)f_i(-D)\,;\ n=0,1,2,\ldots, \tag{2.36a}$$

where

$$\varepsilon_{ijn} = \frac{1}{D}\int_{-D}^{0} f_i(z)f_j(z)f_n(z)dz. \tag{2.36b}$$

In general, the modes are coupled because they interact with the bottom and with each other so energy can leak from one mode to another. But in linear theory and in the case of a flat bottom, the modes are decoupled.

2.4 A Mean State

Let us now introduce a depth-dependent weak mean current $\underline{\bar{v}}(z)$. By "weak", we mean

$$|\underline{\bar{v}}| << U, \tag{2.37}$$

so that $\underline{\bar{v}}$ is small enough to disallow dynamical instabilities. The mean current can also be decomposed into normal modes:

$$\underline{\bar{v}} = [\sum_n \bar{u}_n f_n(z) \ , \quad \sum_n \bar{v}_n f_n(z)], \tag{2.38}$$

where $\bar{u}_n$ and $\bar{v}_n$ are the constant modal amplitudes of the eastward and northward mean currents, respectively, in the nth mode. In general, the kinetic energy of the lower modes dominates, so that the mean current can be parameterized by only a few modal amplitudes, and only these modal amplitudes appear in the horizontal-structure equations to represent the effects of the mean current on wave propagation. From the geostrophic relation we know that the associated mean variation of pressure is

$$\bar{p}' = \sum_n \bar{p}'_n f_n(z), \tag{2.39a}$$

with

$$\bar{p}'_n(x,y) = \rho^* f_0(-\bar{u}_n y + \bar{v}_n x). \tag{2.39b}$$

Of course $\bar{p}'$ must satisfy the time-independent quasigeostrophic potential vorticity equation (2.15), implying that the mean modal-amplitude function $\bar{p}'_n$ must satisfy the corresponding horizontal equation (2.36a). Note that there are mean variations of ρ, η and δc as well, and a zonal mean current over a flat bottom is always a possible mean state.

Let us denote the modal amplitude function of the fluctuating pressure in the nth mode by π_n such that

$$\bar{p}'_n = \bar{p}'_n(x,y) + \pi_n(x,y,t). \tag{2.40}$$

It follows that (2.36a) can be written as

$$\left[\frac{\partial}{\partial t}(\nabla_H^2 - \lambda_n) + \beta\frac{\partial}{\partial x}\right]\pi_n + \sum_{ij} \varepsilon_{ijn}\left(\bar{u}_i\frac{\partial}{\partial x} + \bar{v}_i\frac{\partial}{\partial y}\right)(\nabla_H^2 - \lambda_j + \lambda_i)\pi_j$$
$$+ \frac{1}{\rho^* f_0}\sum_{ij} \varepsilon_{ijn} J[\pi_i, (\nabla_H^2 - \lambda_j)\pi_j] = -\frac{f_0 f}{D}(-D)\sum_i f_i(-D) J(\pi_i, b);$$
$$n = 0, 1, 2, \ldots. \tag{2.41}$$

In the following analyses of wave propagation, we restrict the bottom slopes $b_x = \partial b/\partial x$ and $b_y = \partial b/\partial y$ to be constants. This is the same as requiring the nonlinear terms in x and y of the Taylor series expansion of b about the origin to be of order $\varepsilon^2 H/L$ in distances of order L. In addition, we require small slopes such that

$$\left|b_x\right| \text{ and } \left|b_y\right| \ll \varepsilon H/L, \tag{2.42}$$

for preventing the existence of bottom-trapped waves.

2.5 Dispersive Primary Waves

Previously, McWilliams and Flierl (1975) have shown that over 90 percent of the kinetic energy in MODE was contained in two empirical orthogonal vertical modes that closely resemble the barotropic and the first baroclinic modes of Rossby waves. Richman et al. (1977) have shown that about 90 percent of the potential energy, again in MODE, was contained in the first three baroclinic modes, with 65 percent of the energy being contained in the first mode alone. Moreover, by decomposing the CTD profiles obtained in the tomographic experiment into the normal modes, we have consistently found that the potential energy of the first mode dominates (Ch. 3, Sec. 3.3). Therefore, without discarding the major features of mesoscale perturbations, we can set $\pi_n=0$ for $n>1$ in the horizontal equations (2.36a). The equations consisting only of the lowest two modes can be written as

$$L_0(\pi_0) = -C_0(\pi_1) - \frac{1}{\rho^* f_0}\left[J(\pi_0,\nabla_H^2\pi_0)+J(\pi_1,\nabla_H^2\pi_1)\right] \qquad (2.43a)$$

and

$$L_1(\pi_1) = -C_1(\pi_0) - \frac{1}{\rho^* f_0}\left[\varepsilon_{111}J(\pi_1,\nabla_H^2\pi_1)+J(\pi_1,\nabla_H^2\pi_0)\right. + \left.\frac{1}{\rho^* f_0}J[\pi_0,(\nabla_H^2-\lambda_1)\pi_1]\right. \qquad (2.43b)$$

where

$$L_0 = \frac{\partial}{\partial t}\nabla_H^2 + \beta\frac{\partial}{\partial x} + (\bar{u}_0\frac{\partial}{\partial x} + \bar{v}_0\frac{\partial}{\partial y})\nabla_H^2 + \frac{f_0}{D}(b_y\frac{\partial}{\partial x} - b_x\frac{\partial}{\partial y}), \qquad (2.43c)$$

$$L_1 = \frac{\partial}{\partial t}(\nabla_H^2 - \lambda_1) + \beta\frac{\partial}{\partial x} + (\bar{u}_0\frac{\partial}{\partial x} + \bar{v}_0\frac{\partial}{\partial y})(\nabla_H^2 - \lambda_1) + \varepsilon_{111}(\bar{u}_1\frac{\partial}{\partial x} + \bar{v}_1\frac{\partial}{\partial y})\nabla_H^2 + \frac{f_0 f_1^2}{D}(-D)(b_y\frac{\partial}{\partial x} - b_x\frac{\partial}{\partial y}) \qquad (2.43d)$$

and

$$C_n = (\bar{u}_1\frac{\partial}{\partial x} + \bar{v}_1\frac{\partial}{\partial y})(\nabla_H^2 + \lambda_n) + \frac{f_0}{D} f_1(-D)(b_y\frac{\partial}{\partial x} - b_x\frac{\partial}{\partial y}); \; n=0,1, \qquad (2.43e)$$

are linear operators (note that $\lambda_0 \sim 0$). Before seeking the wave solutions for π_0 and π_1, we make the following observations from (2.43): (1) modes are linearly coupled as denoted by $C_n(\pi_m)$ because the fluid motions interact with the mean current and the bottom slopes, and (2) just like the mean current, the current associated with a wave can advect the vorticity of other waves as denoted by the Jacobians, hence creating nonlinear effects.

The advection of vertical planetary vorticity south to north, which is proportional to the largest term $\beta\partial\pi_n/\partial x$ in $L_n(\pi_n)$, is responsible for the propagations of planetary waves. Whether the linearization for the wave motions is valid or not depends on the smallness of the ratio, ν, of the magnitude of nonlinear terms to

the magnitude of $\beta \partial \pi_n / \partial x$. Qualitatively, the nonlinear terms are of order $U^2 \rho^* f_0 / L^2$ and $\beta \partial \pi_n / \partial x$ is of order $\beta \rho^* f_0 U$. Thus we obtain, approximately,

$$\nu \sim U / L^2 \beta. \tag{2.44}$$

By using the typical values of U, L and β, we obtain $\nu \sim 0.25$. This is not a small value when compared to unity so that nonlinear effects could be important. However, quantitatively, ν can be much smaller depending on the wavenumbers of the interacting waves. The quantitative estimation is defered to Sec. 2.6.1. Let us assume for the moment that $\nu << 1$. By the assumptions of small bottom slopes and weak mean current, we know that the ratio of the magnitudes of $C_n(\pi_m)$ to $L_n(\pi_n)$ is much smaller than unity, and for convenience, let us assume that this ratio is also of order ν so that we can construct the solution for π_n as a perturbation series of powers of ν such that

$$\pi_n = \pi_n^{(0)} + \pi_n^{(1)} + \cdots\cdots \tag{2.45}$$

with $\pi_n^{(i+1)} / \pi_n^{(i)} \sim \nu$. We will call the zeroth-order solution $\pi_n^{(0)}$ and the first order correction $\pi_n^{(1)}$ the primary and secondary perturbations, respectively.

In the zeroth-order approximation, the horizontal equations (2.43a, b) are linearized and decoupled:

$$L_n(\pi_n^{(0)}) = 0 \quad ;n=0,1. \tag{2.46}$$

2.5.1 Dispersion And Phase Velocity

Equation (2.46) admits a free-wave solution. The properties of these waves can be investigated using a triple Fourier transform. Let the complex spectrum (or the Fourier transform) of the modal-amplitude function $\pi_n(x,y,t)$ be $\phi_n(k_n,l_n,\sigma_n)$ such that

$$\pi_n^{(0)}(x,y,t) = \iiint \phi_n(k_n,l_n,\sigma_n) e^{i(k_n x + l_n y - \sigma_n t)} dk_n dl_n d\sigma_n. \tag{2.47}$$

The spectrum shows how the pressure in the nth mode is distributed in the wavenumber-frequency domain, the amplitude of each individual wave being infinitesimal in a continuous spectrum. In the case of a discrete wave in the nth mode with wavenumber vector (k,l) and frequency σ, $|\phi_n|$ would consist of two impulses with equal amplitudes located at $\pm(k,l,\sigma)$. The area under them is the amplitude of the wave.

By Fourier transforming (2.46) and then cancelling ϕ_n, we find

that the waves in the nth mode (n=0,1) must satisfy the dispersion relation

$$\Delta_n(k_n,l_n,\sigma_n) = 0 \tag{2.48a}$$

where

$$\Delta_n(k_n,l_n,\sigma_n) \equiv (k_n^2+l_n^2+\lambda_n)(\sigma_n+\delta\sigma_n)+k_n(\beta+\delta\beta_n), \tag{2.48b}$$

$$\delta\sigma_0 = -(\bar{u}_0 k_0+\bar{v}_0 l_0), \tag{2.48b}$$

$$\delta\sigma_1 = -[k_1(\bar{u}_0+\varepsilon_{111}\frac{k_1^2+l_1^2}{k_1^2+l_1^2+\lambda_1}\bar{u}_1)+l_1(\bar{v}_0+\varepsilon_{111}\frac{k_1^2+l_1^2}{k_1^2+l_1^2+\lambda_1}\bar{v}_1] \tag{2.48c}$$

and

$$\delta\beta_n = \frac{f_0}{D} f_n^2(-D)(b_y-\frac{l_n}{k_n} b_x). \tag{2.48d}$$

By rearranging, we get

$$(\sigma_n+\delta\sigma_n) = \frac{-k_n(\beta+\delta\beta_n)}{k_n^2+l_n^2+\lambda_n} \tag{2.49}$$

As expected, the mean current causes Doppler effects given by $\delta\sigma_n$'s, which vanish when there is no mean current. It is seen that the propagation of barotropic waves are not affected by mean baroclinic currents, and the contributions to Doppler shifts from mean baroclinic currents to the wavefrequencies are minor for baroclinic waves with wavelengths much longer than the radius of deformation $\lambda_1^{-1/2}$, (that is for waves with $k_1^2+l_1^2<<\lambda_1$).

It is the small latitudinal variation of the coriolis parameter (or the β-effect) that allows the propagation of waves with subinertial frequencies by changing the relative vorticity. However, the β-effect on wavefrequencies can be modified by $\delta\beta_n$ in the presence of bottom slopes. This is so because the slopes modify the vertical velocity and hence change the relative vorticity also (see (2.10a) and (2.11)). The modification of frequencies caused by the longitudinal bottom slope b_x is small when waves are propagating zonally, that is when $l_n/k_n<<1$. The β-effect is enhanced or reduced depending on the direction of rising (or falling) topography and the direction of wave propagation. Because the energy of baroclinic waves is trapped more in the upper water column than that of the barotropic waves, baroclinic waves are less affected by the slopes (note that $\left|f_1(-D)\right| <f_0(-D)=1$ in (2.48d)).

The phase-velocity vector of a wave in the nth mode is

$$\underline{c}_n = \left(\frac{\sigma_n}{k_n}, \frac{\sigma_n}{l_n}\right) = \left[-\left(\frac{\beta+\delta\beta_n}{k_n^2+l_n^2+\lambda_n} + \frac{\delta\sigma_n}{k_n}\right), -\left(\frac{k_n}{l_n}\right)\frac{\beta+\delta\beta_n}{k_n+l_n+\lambda_n} - \frac{\delta\sigma_n}{l_n}\right]. \quad (2.50)$$

The east component of $\underline{c}_n$ is almost always negative because of small slopes and weak mean current that usually imply $\beta > |\delta\beta_n|$ and $\beta/(k_n^2+l_n^2+\lambda) > |\delta\sigma_n/k_n|$. This feature of westward phase propagation is generally observed in experiments.

By rearranging (2.49), we get

$$\left(k_n + \frac{1}{2}\frac{\beta+\delta\beta_n}{\sigma_n+\delta\sigma_n}\right)^2 + l_n^2 = \left(\frac{1}{2}\frac{\beta+\delta\beta_n}{\sigma_n+\delta\sigma_n}\right)^2 - \lambda_n. \quad (2.51)$$

Since k_n and l_n are real for propagating waves, the R.H.S. of (2.51) must be positive such that

$$\sigma_n \leq \frac{\lambda_n^{-1/2}}{2}(\beta+\delta\beta_n) - \delta\sigma_n, \quad (2.52)$$

implying that there is a frequency limit for wave propagation. In general, the upper cutoff frequency depends on mode number, wavenumber, mean current, and bottom slope. Because $\lambda_0^{-1/2} >> \lambda_1^{-1/2}$, the cutoff frequencies for baroclinic waves are much smaller than those of barotropic waves.

The representation of $\pi_n^{(0)}$ by a continuous sum of its wavenumber-frequency components distributed in the three dimensional wavenumber-frequency spectrum $\phi_n(k_n, l_n, \sigma_n)$ is adequate but nolonger necessary due to the dispersion relation. A full description of the fluctuating field can be provided just as well by the simplier two dimensional wavenumber-spectrum $\phi_n(k_n, l_n)$ such that

$$\pi_n^{(0)} = \iint \phi_n(k_n, l_n) e^{i(k_n x + l_n y - \sigma_n t)} \, dk_n \, dl_n . \qquad (2.53)$$

2.5.2 Narrowband Processes And Group Velocity

For fluctuations due to narrowband processes in the wave-number spectrum, $|\phi_n(k_n,l_n)|$ contains pulses of finite width and $\pi_n^{(0)}$ can be represented by a sum of modulated waves. With a total number of N_n pairs of pulses in $|\phi_n|$ and with the ith pair being located at $\pm(k_{ni},l_{ni})$, we can write

$$\pi_n^{(0)}(x,y,t) = \sum_{i=1}^{N_n} a_{ni}(x,y,t)\cos(k_{ni}x+l_{ni}y-\sigma_{ni}t+\gamma_{ni}), \quad (2.54)$$

where each modulating amplitude (or envelope) a_{ni} is slowly varying in space and time as compared to its carrier which has a phase constant γ_{ni} and a frequency σ_{ni} that satisfies the dispersion relation. The slowly varying nature of a_{ni} in x and y is implied directly by the narrowband processes in the wavenumber spectrum; the slowly varying nature in t can be verified by investigating the group velocity.

While the phases of the carrier waves are propagating with the phase velocities, the phases of their envelopes are propagating with the corresponding group velocities. The group-velocity vector can be evaluated by

$$\underline{v}_{gn}(k_n,l_n) = \left(\frac{\partial\sigma_n}{\partial k_n}, \frac{\partial\sigma_n}{\partial l_n}\right). \quad (2.55)$$

The result (which is not shown here) is a complicated vector function of k_n and l_n, indicating that in general the modulating envelopes can propagate in any direction and that the group speeds are much smaller than the corresponding phase speeds. Since a_{ni} is varying very slowly in time and space, $\pi_n^{(0)}$ can be approximated locally by a sum of discrete waves with constant amplitudes a_{ni}, where a_{ni} is equal to the area under the ith pair of pulses in $|\phi_n|$.

2.6 Mode Couplings and Nonlinear Interactions

Since the coupling and nonlinear terms in the horizontal equations are not identically zero but finite, intermodal wave forcing and nonlinear wave-wave interactions must occur during wave propagation. In order to investigate coupling and nonlinear effects, we must proceed to the next order in ν.

To order ν, we have

$$L_0(\pi^{(1)}) = -C_0(\pi_1^{(0)}) - \frac{1}{\rho^* f_0}[J(\pi_0^{(0)}, \nabla_H^2 \pi_0^{(0)}) + J(\pi_1^{(0)}, \nabla_H^2 \pi_1^{(0)})] \tag{2.56a}$$

and

$$L_1(\pi_1^{(1)}) = -C_1(\pi_1^{(0)}) - \frac{1}{\rho^* f_0}[\varepsilon_{111} J(\pi_1^{(0)}, \nabla_H^2 \pi_1^{(0)}) + J(\pi_1^{(0)}, \nabla_H^2 \pi_0^{(0)})] + \frac{1}{\rho^* f_0} J[\pi_0^{(0)}, (\nabla_H^2 - \lambda_1)\pi_1^{(0)}]. \tag{2.56b}$$

It is seen that the zeroth-order solutions $\pi_n^{(0)}$ are now the forcing mechanisms for the first-order terms $\pi_n^{(1)}$. This implies that secondary waves of smaller amplitude can be generated by the primary waves through their nonlinear interactions and the linear couplings. If some of the forced (secondary) waves are at resonance, that is their wavenumbers and frequencies also satisfy the dispersion relation (a secular effect), their amplitudes will not remain small but will grow, and at some time will become dominant among all the forced waves.

Before going into the subject of forced and resonant waves in more detail, we will first come back to the issue of whether the effects of the nonlinear interactions of primary waves are small or not. The issue is important because the validity of the asymptotic solution constructed as a perturbation series in powers of ν depends on the smallness of ν.

2.6.1 Magnitudes Of The Nonlinear Terms

It was mentioned earlier that ν is of order $U/L^2\beta \sim 0.25$ and is not qualitatively small. But quantitatively, it can be smaller depending on the wavenumbers of the interacting primary waves. This fact will be demonstrated in this section.

There are three cases that we need to consider. They are the interactions between (1) two barotropic waves, (2) two baroclinic waves and (3) one barotropic and one baroclinic wave. We do not need to consider cases for more than two waves because each combination of two can be considered seperately. When we say a wave, it could imply either one wave that is associated with a discrete (or narrowband) spectrum or one infinitesimal group of waves centered at some wavenumber in a continuous spectrum.

In cases (1) and (2), the only nonvanishing nonlinear term is proportional to $J(\pi_n^{(0)}, \nabla_H^2 \pi_n^{(0)})$ with n=0 and n=1 for the first and second cases, repectively, and

$$\pi_n^{(0)} = a_{n1}\cos\theta_{n1} + a_{n2}\cos\theta_{n2}, \tag{2.57}$$

where a_{ni} is the amplitude and $\theta_{ni}=k_{ni}x+l_{ni}y-\sigma_{ni}t+\gamma_{ni}$ is the phase of the ith wave; i=1, 2. Note that $a_{ni}=2\iint b_n(k_{ni},l_{ni})dk_n dl_n$ in the case of a continuous spectrum. The nonvanishing Jacobian term can be cast as

$$J(\pi_n^{(0)},\nabla_H^2\pi_n^{(0)})=J(a_{n1}\cos\theta_{n1},\nabla_H^2 a_{n1}\cos\theta_{n1})+J(a_{n1}\cos\theta_{n1},\nabla_H^2 a_{n2}\cos\theta_{n2})$$
$$+J(a_{n2}\cos\theta_{n2},\nabla_H^2 a_{n1}\cos\theta_{n1})+J(a_{n2}\cos\theta_{n2},\nabla_H^2 a_{n2}\cos\theta_{n2}). \quad (2.58)$$

But since

$$J(a_{ni}\cos\theta_{ni},\nabla_H^2 a_{ni}\cos\theta_{ni}) = 0, \quad (2.59)$$

(2.58) becomes, after performing the Jacobian operation,

$$J(\pi_n^{(0)},\nabla_H^2\pi_n^{(0)}) = \frac{1}{2}a_{n1}a_{n2}[(k_{n1}^2+l_{n1}^2)-(k_{n2}^2+l_{n2}^2)](k_{n2}l_{n1}-k_{n1}l_{n2})$$
$$\times[\cos(\theta_{n1}+\theta_{n2})-\cos(\theta_{n1}-\theta_{n2})]. \quad (2.60)$$

It is seen that the magnitude of the nonlinear term depends on the difference of the squared magnitudes of the wavenumber vectors and the difference of the directions of propagation; the smaller the differences are, the smaller the nonlinear effects. In the limit when the waves have either the same wavelength or the same direction of propagation, there cannot be any nonlinear interactions, and the waves will be an exact solution to the quasigeostrophic vorticity equation. From (2.59), we notice that a single wave is always an exact solution.

In case (3), with

$$\pi_0^{(0)} = a_{01}\cos\theta_{01} \tag{2.61a}$$

and

$$\pi_1^{(0)} = a_{11}\cos\theta_{11}, \tag{2.61b}$$

the sum of the nonvanishing nonlinear terms is proportional to

$$J[\pi_0^{(0)}, (\nabla_H^2-\lambda_1)\pi_1^{(0)}]+J(\pi_1^{(0)}, \nabla_H^2\pi_0^{(0)}) = \frac{1}{2} a_{01}a_{11}(k_{11}l_{01}-l_{11}k_{01}) \times[(k_{11}^2+l_{11}^2+\lambda_1)-(k_{01}^2+l_{01}^2)][\cos(\theta_{01}+\theta_{11})+\cos(\theta_{01}-\theta_{11})] \tag{2.62}$$

It is found here that the magnitude of the nonlinear term again depends on the difference of the directions of propagation, and also depends on the difference of $(k_{11}^2+l_{11}^2)+\lambda_1$ and $(k_{01}^2+l_{01}^2)$. Similarly, the smaller the differences are, the smaller the nonlinear effects. There would not be any nonlinear interactions if either the waves of different modes were propagating in the same direction or the difference of the squared magnitudes of the barotropic and the baroclinic wavenumber vectors were exactly λ_1.

In conclusion, inorder for the (asymptotic) theory of weak wave-wave interactions, which predicts the propagation of forced waves and resonant interactions, to be applicable, the wavenumbers of the primary waves must be so arranged that they make $\nu \ll 1$.

2.6.2 Forced Secondary Waves

In this and the following sections, we will discuss only forced and resonant waves that correspond to two primary baroclinic waves. The other two cases can be investigated by a similar procedure; their results are summarized in tables (2.2) and (2.4) without futher discussion. For the case of more than two primary waves, it is obvious that the forced solutions due to each primary wave in the linear coupling terms and each combination of two primary waves in the nonlinear terms can be summed together to give the total solution.

Secondary perturbations are driven by the primary dispersive waves. For two existing primary baroclinic waves such that

$$\pi_n^{(0)} = a_{11}\cos\theta_{11} + a_{12}\cos\theta_{12}, \qquad (2.63)$$

the governing equation (2.56) for the secondary perturbations becomes

$$L_0(\pi_0^{(1)}) = -C_0(a_{11}\cos\theta_{11}) - C_0(a_{12}\cos\theta_{12}) - \frac{a_{11}a_{12}}{2\rho^* f_0}[(k_{11}^2+l_{11}^2)-(k_{12}^2+l_{12}^2)] \times (k_{12}l_{11}-k_{11}l_{12})[\cos(\theta_{11}+\theta_{12})-\cos(\theta_{11}-\theta_{12})] \qquad (2.64a)$$

and

$$L_1(\pi_1^{(1)}) = -\frac{\varepsilon_{111}a_{11}a_{12}}{2\rho^* f_0}[(k_{11}^2+l_{11}^2)-(k_{12}^2+l_{12}^2)](k_{12}l_{11}-l_{11}l_{12}) \times [\cos(\theta_{11}+\theta_{12})-\cos(\theta_{11}-\theta_{12})]. \qquad (2.64b)$$

Note that beside secondary baroclinic perturbations, secondary barotropic perturbations are also possible due to mode coupling. (Note also that mode coupling can modify the frequencies of the primary waves.) While the forcing produced by linear coupling has components that oscillate with the same primary wavenumbers and frequencies, the forcings produced by nonlinear interactions have components that oscillate with the sums and differences of the primary wavenumbers and frequencies.

The equations for the $\pi_n^{(1)}$'s are linear but nonhomogeneous, containing simple harmonic forcing functions in space and time; therefore the steady-state solutions have the same harmonic forms as the forcing functions. By expecting a phase lead or lag of 90°, we can write down the solution as

$$\pi_0^{(1)} = \sum_{i=1}^{4} b_{0i} \sin\alpha_{0i} \tag{2.65a}$$

and

$$\pi_1^{(1)} = \sum_{i=1}^{2} b_{1i} \sin\alpha_{1i}, \tag{2.65b}$$

where $\alpha_{0i}=\alpha_{1i}=\theta_{11}\pm\theta_{12}$ for $i=1,2$, $\alpha_{03}=\theta_{11}$ and $\alpha_{04}=\theta_{12}$. With the use of (2.64), the wave amplitudes b_{ni} are evaluated as

$$b_{01} = \frac{-Q}{\Delta_0(k_{11}+k_{12}, l_{11}+l_{12}, \sigma_{11}+\sigma_{12})}, \tag{2.66a}$$

$$b_{02} = \frac{-Q}{\Delta_0(k_{11}-k_{12}, l_{11}-l_{12}, \sigma_{11}-\sigma_{12})}, \tag{2.66b}$$

$$b_{03} = \frac{-[(\bar{u}_1 k_{11}+\bar{v}_1 l_{11})(k_{11}^2+l_{11}^2)-(f_0/D)f_1(-D)(b_y k_{11}-b_x l_{11})]}{\Delta_0(k_{11}, l_{11}, \sigma_{11})}, \tag{2.66c}$$

$$b_{04} = \frac{-[(\bar{u}_1 k_{12}+\bar{v}_1 l_{12})(k_{12}^2+l_{12}^2)-(f_0/D)f_1(-D)(b_y k_{12}-b_x l_{12})]}{\Delta_0(k_{12}, l_{12}, \sigma_{12})}, \tag{2.66d}$$

$$b_{11} = \frac{-\varepsilon_{111} Q}{\Delta_1(k_{11}+k_{12}, l_{11}+l_{12}, \sigma_{11}-\sigma_{12})} \tag{2.66e}$$

and

$$b_{12} = \frac{-\varepsilon_{111} Q}{\Delta_1(k_{11}-k_{12}, l_{11}-l_{12}, \sigma_{11}-\sigma_{12})}, \tag{2.66f}$$

where

$$Q = \frac{a_{11}a_{12}}{2\rho^* f_0}[(k_{11}^2+l_{11}^2)-(k_{12}^2+l_{12}^2)](k_{12}l_{11}-k_{11}l_{12}). \tag{2.66g}$$

The secondary perturbations consist of forced waves. In this case, there are four forced barotropic waves and two forced barocl inc waves. Their amplitudes are a factor ν smaller than those of the primary waves except at resonance. Moreover, they need continuous forcing to exist, that is the primary waves must be quite permanent for the forced secondary waves to exist.

2.6.3 Resonant Secondary Waves

When a forced wave of the nth mode with wavenumbers (k_{nf},l_{nf}) and frequency σ_{nf} satisfies the dispersion relation

$$\Delta_n(k_{nf},l_{nf},\sigma_{nf}) = 0, \tag{2.67}$$

resonance occurs and (2.67) is the resonance condition. At resonance, the expressions shown in the last section for wave amplitudes are nolonger valid because the denominator is identically zero and the resonant wave amplitude is growing linearly in time.

The two forced barotropic waves with phases θ_{11} and θ_{12} cannot be resonant because the wavenumbers and frequencies that satisfy the baroclinic dispersion relation can never, at the same time, satisfy the barotropic resonance condition due to the form of the dispersion relation. However, the other four with $(k_{nf},l_{nf})=(k_{11}\pm k_{12},l_{11}\pm l_{12})$, $\sigma_{nf}=\sigma_{11}\pm\sigma_{12}$ and n=1,2

are possible resonant waves. Suppose resonance occurs at n=1. Then at the sums of the wavenumbers and frequencies, there will be a resonant baroclinic wave having the form $b_{11}(t)\cos(\theta_{11}+\theta_{12})$. With the use of (2.64b), we find that the growth rate of the amplitude is

$$\frac{db_{11}}{dt} = \frac{\varepsilon_{111}Q}{(k_{nf}^2+l_{nf}^2+\lambda_1)} \quad . \tag{2.68}$$

The growth rates evaluated at resonance for the other three possibilities are shown in table (2.3).

It is interesting to point out that the growth rates do not depend on the mean current and the bottom topography, but are proportional to the magnitudes of the corresponding nonlinear terms.

Table 2.2

Interaction Between 2 Oth-Mode Primary Waves

mode	primary wave		secondary wave		resonance	growth
	amplitude	phase	amplitude	phase		rate
0th	a_{01}	θ_{01}	b_{01}	$\theta_{01}+\theta_{02}$	possible	g_{01}
	a_{02}	θ_{02}	b_{02}	$\theta_{01}-\theta_{02}$	possible	g_{02}
1st			b_{11}	θ_{01}	no	
			b_{12}	θ_{02}	no	

$$\theta_{\substack{01\\02}}=k_{\substack{01\\02}}x+l_{\substack{01\\02}}y-\sigma_{\substack{01\\02}}t+\gamma_{\substack{01\\02}}$$

$$Q=a_{01}a_{02}[(k_{01}^2+l_{01}^2)-(k_{02}^2+l_{02}^2)](k_{02}l_{01}-k_{01}l_{02})/2\rho f_0$$

$$b_{\substack{01\\02}}=\frac{-Q}{\Delta_0(k_{01}\pm k_{02},l_{01}\pm l_{02},\sigma_{01}\pm\sigma_{02})} \qquad g_{\substack{01\\02}}=\frac{Q}{(k_{01}\pm k_{02})^2+(l_{01}\pm l_{02})^2}$$

$$b_{\substack{11\\12}}=\frac{-[(\bar{u}_1k_{01}^2+\bar{v}_1k_{01}^2)(k_{01}^2+l_{01}^2)-(f_0/D)f_1(-D)(b_yk_{01}-b_xl_{01})}{\Delta_1(k_{01},l_{01},\sigma_{01})}$$

Table 2.3

Interaction Between 2 1st-Mode Primary Waves

mode	primary wave amplitude	primary wave phase	secondary wave amplitude	secondary wave phase	resonance	growth rate
0th			b_{01}	$\theta_{11}+\theta_{12}$	possible	g_{01}
			b_{02}	$\theta_{11}-\theta_{12}$	possible	g_{02}
			b_{03}	θ_{11}	no	
			b_{04}	θ_{12}	no	
1st	a_{11}	θ_{11}	b_{11}	$\theta_{11}+\theta_{12}$	possible	g_{11}
	a_{12}	θ_{12}	b_{12}	$\theta_{11}-\theta_{12}$	possible	g_{12}

$$\theta_{\substack{11\\12}}=k_{\substack{11\\12}}x+l_{\substack{11\\12}}y-\sigma_{\substack{11\\12}}t+\gamma_{\substack{11\\12}}$$

$$Q=a_{11}a_{12}[(k_{11}^2+l_{11}^2)-(k_{12}^2+l_{12}^2)](k_{12}l_{11}-k_{11}l_{12})/2\rho^*f_0$$

$$b_{\substack{01\\02}}=\frac{-Q}{\Delta_0(k_{11}\pm k_{12},l_{11}\pm l_{12},\sigma_{11}\pm\sigma_{12})} \qquad g_{\substack{01\\02}}=\frac{Q}{(k_{11}\pm k_{12})^2+(l_{11}\pm l_{12})^2}$$

$$b_{\substack{03\\04}}=\frac{-[(\bar{u}_1k_{\substack{11\\12}}^2+\bar{v}_1k_{\substack{11\\12}}^2)(k_{\substack{11\\12}}^2+l_{\substack{11\\12}}^2)-(f_1/D)(b_yk_{\substack{11\\12}}-b_xl_{\substack{11\\12}})}{\Delta_0(k_{\substack{11\\12}},l_{\substack{11\\12}},\sigma_{\substack{11\\12}})}$$

$$b_{\substack{11\\12}}=\frac{-\varepsilon_{111}Q}{\Delta_1(k_{11}\pm k_{12},l_{11}\pm l_{12},\sigma_{11}\pm\sigma_{12})} \qquad g_{\substack{11\\12}}=\frac{\varepsilon_{111}Q}{(k_{11}\pm k_{12})^2+(l_{11}\pm l_{12})^2+\lambda_1}$$

Table 2.4

Interaction Between 1 0th-Mode and 1 1st-Mode Primary Waves

mode	primary wave		secondary wave		resonance	growth
	amplitude	phase	amplitude	phase		rate
0th	a_{01}	θ_{01}	b_{01}	θ_{11}	no	
1st	a_{11}	θ_{11}	b_{11}	$\theta_{01}+\theta_{11}$	possible	g_{11}
			b_{12}	$\theta_{01}-\theta_{11}$	possible	g_{12}
			b_{13}	θ_{01}	no	

$$\theta_{\substack{01\\11}}=k_{\substack{01\\11}}x+l_{\substack{01\\11}}y-\sigma_{\substack{01\\11}}t+\gamma_{\substack{01\\11}}$$

$$Q=a_{01}a_{11}[(k_{01}^2+l_{01}^2)-(k_{11}^2+l_{11}^2+\lambda_1^2)](k_{11}l_{01}-k_{01}l_{11})/2\rho f_0$$

$$b_{01}=\frac{-[(\bar{u}_1k_{11}^2+\bar{v}_1k_{11}^2)(k_{11}^2+l_{11}^2)-(f_0/D)(b_yk_{11}-b_xl_{11})}{\Delta_0(k_{11},l_{11},\sigma_{11})}$$

$$b_{\substack{11\\12}}=\frac{-\varepsilon_{111}Q}{\Delta_1(k_{11}\pm k_{12},l_{11}\pm l_{12},\sigma_{11}\pm\sigma_{12})} \qquad g_{\substack{11\\12}}=\frac{\varepsilon_{111}Q}{(k_{11}\pm k_{12})^2+(l_{11}\pm l_{12})^2+\lambda_1}$$

$$b_{01}=\frac{-[(\bar{u}_1k_{01}^2+\bar{v}_1k_{01}^2)(k_{01}^2+l_{01}^2)-(f_0/D)f_1(-D)(b_yk_{01}-b_xl_{01})}{\Delta_0(k_{01},l_{01},\sigma_{01})}$$

CHAPTER 3

THE FORWARD PROBLEM: RELATING OBSERVATIONS TO WAVE PARAMETERS

While the baroclinic planetary waves produce significant changes in both the horizontal-current and vertical-displacement fields (i.e., temperature field), the barotropic planetary waves produce only significant changes in the horizontal-current field and very little vertical displacements. Thus, the baroclinic waves are observable through temperature measurements alone but the observations of both types of waves must be accomplished with combined measurements of current and temperature.

In our investigation of the existence and dynamics of planetary waves, we used the different types of temperature measurements obtained in the 1981 Ocean Acoustic Tomography Experiment. Data were provided by The Ocean Tomography Group. Although current measurements were also available, they were not used in the study. The current measurements lack spatial resolution since current meters were mounted on two enviromental moorings only (but we have used the temperature records from those moorings). Thus, we are limited to the detection of the baroclinic waves only. Three types of temperature measurements were made. They are the in-situ profiles, the point measurements, and the integral measurements (i.e., the acoustic travel times), obtained from the CTD surveys, the moored temperature recorders and sensors, and the acoustic tomographic array, respectively.

In order to extract information on the baroclinic waves from the temperature measurements, the forward problem of how the temperature field and its measurements are affected by the evolution of the waves must first be resolved. This is done in this chapter in conjunction with the last one (Ch. 2). In the last chapter, we studied the theory of planetary waves by reviewing the literature. We saw that the space and time behavior of the wave-induced perturbations of sound speed (or equivalently of temperature) are constrained by the modal dispersion relationships and characterized by the wave parameters such as the wavenumbers, wave amplitudes, modal amplitudes of the mean flow, etc. In this chapter, the objective is to develop the model equations that relate the data to the unknown parameters that characterize the wave-induced perturbations and mean-flow induced variations of sound speed. In Ch. 5, the model equations are inverted for the wave parameters. Of course, one can use either the perturbations of sound speed or temperature as the observed dynamical variable in the model equations, for the two variables are intimately related and empirically proportional to each other (Wilson, 1960 and Medwin, 1975). We prefer to use the sound-speed perturbation δc.

We begin in this chapter by giving a brief description of the 1981 Ocean Tomography Experiment. For a detailed description of the experiment, the reader is referred to the Ocean Tomography Group (1982). Next, the empirical relation between temperature and δc and the integral relation between perturbation of acoustic travel time

and δc are discussed. We also discuss the data set actually being used in the model equations for the parameter estimations. The data set was obtained by filtering (daily averaging) the point and integral measurements and compressing the profile measurements. Finally, we present three plausible dynamical models for wave propagation and develop the model equations. The space and time behavior of the wave-induced δc is constrained and characterized differently in the different models. By fitting the different wave-propagation models to the data set, the wave dynamics are then estimated in Ch. 5.

3.1 The Experiment

In the spring of 1981, the Ocean Tomography Group conducted the first field test of a full tomographic system in a 300 km square south-west of Bermuda over a period of 4 months (Ocean Tomography Group, 1982). The goal was to test the practicality of the acoustic inverse scheme of Munk and Wunsch (1979) for monitoring the ocean interior at mesoscale resolution. In order to evaluate the performance of the tomographic system, the region was also measured by traditional techniques during the same period so that a basis for comparison could be provided. The tomographic data was inverted by Cornuelle (1983) and Cornuelle et al. (1985) on a daily basis; the daily tomographic maps he generated compare favorably with the ship-based objective maps. This work demonstrated the great potential of acoustic tomography for adequate and effective large scale monitoring. Here, our chief goal is to investigate the existence and dynamics of planetary waves, and in order to make the best estimates of the wave parameters and wave dynamics, we incorporate all types of temperature measurements in our inversions.

The experimental square was centered at 26°N, 70°w over the Hatteras abyssal plain and just south of the region in which MODE was conducted. The ocean bottom here has a nominal depth of 5400 m and a very small depth variation of 300 m over the square. The tomographic system itself consisted of a horizontal array of 4 sources and 5 receivers moored at a nominal depth of 2000 m

surrounding the square. All the acoustic sources (Si; i=1,2,3,4) were moored at the western boundary, 4 of the receivers (Ri; i=1,2,3,4) were moored at the eastern boundary and the remaining receiver (R5) was moored near the northern boundary of the square. Using the signal processing technique of Spindel (1979), a 224 Hz carrier modulated by a repetition of a maximal length shift register sequence that lasted nearly 3 minutes was transmitted hourly on every third day between each of the source-receiver pairs, and through a form of matched filtering, the multipath travel times of the sequence were estimated. Although the transmissions were intended to last for 4 months, more than half of the receivers had stopped recording data after 80 days into the experiment due to failure of the batteries. The motions of the acoustic moorings were tracked by bottom-mounted acoustic transponders. The tracking was needed to prevent the misinterpretation of the large changes in travel times due to mooring motion as changes due to oceanic perturbations. However, some of the tracking data were missing and hence mooring motions must also be dealt with in the model equations; that is in addition to δc, the uncertainty of the the positions of the sources and receivers must also be modelled. (Cornuelle, 1983 contains a detailed discussion of how to model the mooring motions.)

The horizontal geometry of the tomographic array is shown in Fig. 3.1. Besides the 9 acoustic moorings, 2 environmental moorings, denoted by E1 and E2 in Fig 3.1, were also deployed.

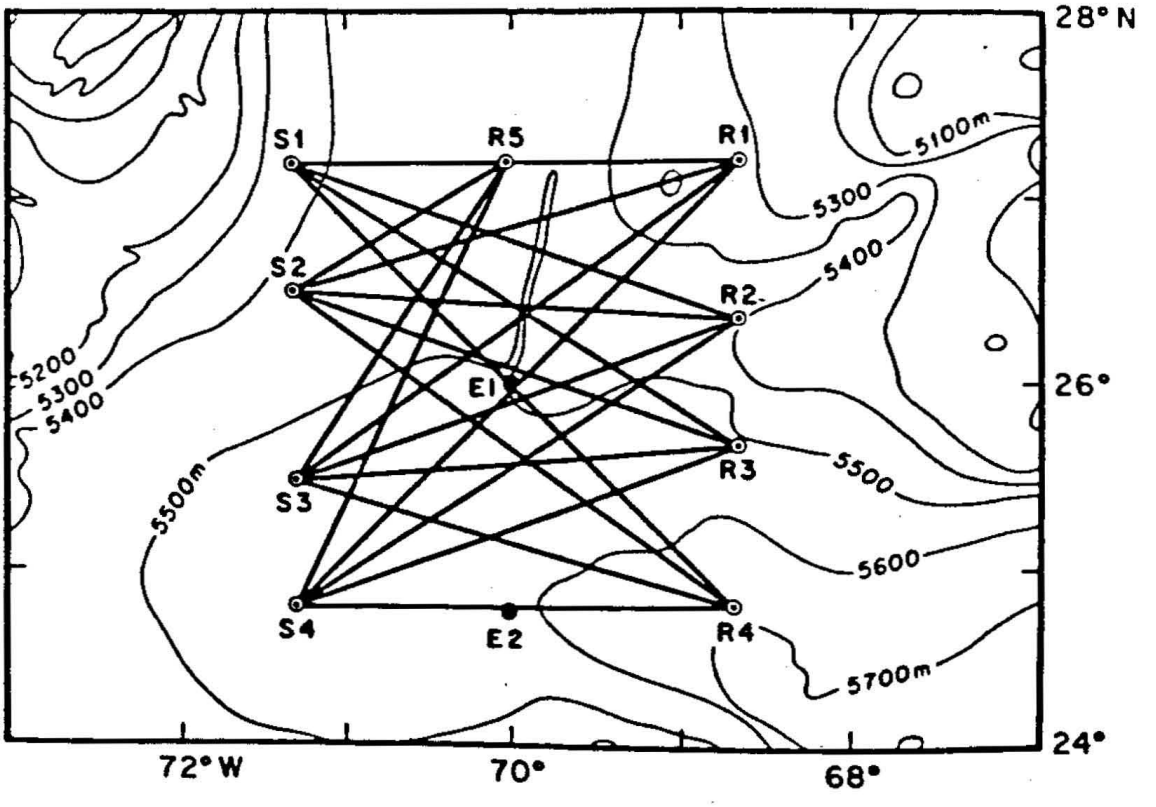

Figure 3.1. The horizontal geometry of the 1981 Ocean Acoustic Tomography Experiment (from Cornuelle et al., 1985), showing 4 source moorings (S1, S2, S3 and S4), 5 receiver moorings (R1, R2, R3, R4 and R5) and 2 enviromental moorings (E1 and E2). The diagram also shows the topography of the experimental region.

Current meters were mounted on the environmental moorings but not on the acoustic moorings. A total of 32 temperature-pressure recorders and temperature sensors were distributed on the moorings and mounted at different depths. However, most of them were not useful for our purpose because they were mounted either in shallow (above 300 m) or deep (below 1600 m) water, where information on the lowest baroclinic-mode planetary waves is hardly obtainable. While the temperature field in the upper layers cannot be described by the lower modes alone and contains strong higher-mode perturbations, the data obtained in the deep zone contain little wave signal (i.e. shows very little variation).

Three CTD surveys in March, May and July and 2 AXBT surveys in April and June were conducted. Each CTD survey lasted 2.5 weeks and had 65 casts distributed evenly over most of the square, but denser in the middle. Each AXBT survey had drops distributed at the same locations as the CTD stations, but such drops are limited to surveying the upper layer of the ocean only, and thus are not useful for our purpose.

3.2 Observations of sound speed perturbations.

3.2.1 Profile and Point Measurements

The speed of sound in water, c, is given by the square root of the ratio of the adiabatic compressibility and density (a derivation of the relation can be found in Clay and Medwin, 1977). As the adiabatic compressibility and density depend on temperature T, salinity S and pressure p (or depth -z) so does c. Empirical formulae for the determination of c from T, S and p or z have been generated by oceanographic acousticians using regression techniques and polynomial least square fittings of laboratory velocimeter measurements of sea water sound speed over large ranges of S, T and p. Some of the well-known and highly accurate formulae are those of Wilson (1960), Medwin (1975) and Lovett (1978); they give almost identical results for the sound speed.

The empirical formulae make it possible to relate CTD surveys to the observations of δc profiles. We prefer the formula of Medwin (1975) for its simplicity; it is given by

$$c = 1449.2+4.6T-0.055T^2+0.00029T^3 + (1.34-0.010T)(S-35)-0.016z, \qquad (3.1)$$

where the physical dimensions of c, T, S and z in the equation are m/s, °C, parts per thousand and m, respectively. CTD casts can be converted to sound-speed profiles by (3.1), and a mean profile $\overline{c}(z)$

can be estimated by averaging all the profiles. Thus, for each CTD cast, a profile measurement of sound-speed perturbation can simply be obtained by

$$\delta c(z) = c(z) - \bar{c}(z). \tag{3.2}$$

A mean temperature profile $\bar{T}(z)$ can be estimated by averaging all the surveyed temperature profiles. By varying c with respect to T in (3.1) and neglecting the salinsity effects, we obtain the empirical relation between δc and temperature perturbation $\delta T = T - \bar{T}$, that is

$$\delta c = 4.6\delta T - 0.11\bar{T}\delta T + 0.00087\bar{T}^2\delta T. \tag{3.3}$$

Using (3.3), time series of the sound-speed perturbation can be obtained from moored time records of temperature.

We have converted all the CTD profiles and temperature time records measured in the experiment to profiles and time series of δc, using (3.1), (3.2) and (3.3), respectively.

3.2.2 Integral Measurements

The description of the acoustic field in a moving medium by an approximate solution using geometrical optics is valid when the changes in pressure, density and entropy of the medium are small over the wavelengths of the sound being transmitted (Blokhintsev, 1956). Such a description is known as ray acoustics and is appropriate for the case of underwater sound transmissions in deep water at relatively high acoustic frequencies, of order 200 Hz and higher. (A frequency of 224 Hz was used in the tomographic system.) The ray solution, that is the geometrical optics approximation, for the acoustic pressure at a frequency ω_a can be cast as

$$p_a(\underline{x},t) = A(x)e^{i\omega_a[c^*\theta(\underline{x})-t]}, \tag{3.4}$$

where c^* is an arbitrary constant reference sound speed, and A is the amplitude and $\omega_a c^*\theta$ is the phase of the time-independent component of p_a. Blockhintsev (1956) has presented a detailed derivation of the differential equations that govern A and θ. The equation for θ is commonly known as the eikonal equation, relating θ to the perturbed sound-speed field $\bar{c}+\delta c$ and the flow field $\underline{v}$ during a transmission by

$$\left|\nabla\theta\right|^2 = (c^*-\underline{v}\cdot\nabla\theta)^2/(\bar{c}+\delta c)^2. \tag{3.5}$$

In general, δc and v vary in both space and time, but they are considered as time invariant in the derivation of (3.5), because they vary on a time scale which is usually much longer than the duration of a transmission so that the ocean can be assumed to be "frozen" momentarily. We will not concern ourselves with the equation for A, which is known as the transport equation, because A is not directly related to the travel-time data. However, it is worth mentioning that the solution for A is important for the identification of multipath arrivals. The interested reader should consult Spiesberger et al. (1980) for ray identifications.

Ugincius (1970) solved the eikonal equation using the method of characteristics and, by proving that the direction of the characteristics and acoustic ray paths coincide, he then obtained the equation for the ray paths from the equation of the characteristics. In a slowly moving and almost stratified medium with $|\underline{v}/c|^2$, $|(dc/dx)/(dc/dz)|$ and $|(dc/dy)/(dc/dz)|$ being much smaller than unity, the ray equation can be approximated by

$$\frac{d}{ds}\left[\frac{c^*}{\bar{c}+\delta c}\frac{d}{ds}(\underline{x}+\underline{\delta x}) - \frac{c^*\underline{v}}{(\bar{c}+\delta c)^2}\right] = 0, \tag{3.6}$$

where s is the arc length along a ray path, $\underline{x}=\underline{x}(s)$ is the nominal trajectory of the ray path in the unperturbed and motionless state and $\underline{\delta x}=\underline{\delta x}(s)$ is the deviation from $\underline{x}(s)$ due to the existence of δc and $\underline{v}$. For the case of mesoscale eddies, $|\underline{v}/c|^2$ is of order

10^{-10} and the ratio of the horizontal to the vertical sound-speed gradients is of order 10^{-5}; thus the ratios are indeed much smaller than unity. The approximate equation (3.6) has the solution of planar rays, that is rays that start out in a vertical plane will always remain in that plane. Munk (1980) considered the effect of horizontal mesoscale sound-speed gradients on horizontal ray bending, but found that the bending is negligible, with the maximum deflection angle being smaller than the horizontal fractional change in the sound speed. By definition, a ray path is a direction of transport of acoustic energy, and the direction is the same as the normal to the wavefront (i.e. $\nabla\theta$) only when the medium is motionless. With fixed locations of acoustic source and receiver, (3.6) is an eigenvalue problem. This implies that depending on the sound-speed profile, sound energy may propagate in more than one discrete direction before reaching the receiver, that is there may be many ray paths that connect a source-receiver pair. Multipath propagation is indeed a prominent feature in the mid-ocean sound channel and the feature is fully exploited by acoustic tomography in attaining vertical resolution in the estimation of the perturbed sound-speed field.

It is indicated in (3.1) that as temperature or pressure increases, so does sound speed. A consequence of the competition between decreasing temperature and increasing pressure with depth, typical at mid latitudes, is the formation of a sound-speed minimum at a depth of about 1 km. This can be seen in Fig. 3.2a in which an

average sound-speed profile in the tomographic region is plotted. The sound-speed minimum (or "axis") of the sound channel traps some sound energy within it. For sound waves that progress forward in either an upward or downward direction, the increase of sound speed tends to refract them back to the axis. The trapped energy propagates along numerous refracted ray paths that sample different vertical sections of the water column and collects information about the perturbed sound-speed field through the accumulated travel-time changes. (Fig. 3.2b shows the geometry of some of the eigen-rays that connected the source S4 and the receiver R3.) In a pulse transmission, the trapped energy in the form of multipath arrivals can be detected over a long distance by a receiver being placed near or at the axis. Thus once the multipath arrivals of each of the source-receiver transmissions in a tomographic array are identified and resolved, they can be used alone or together with other measurements to estimate the perturbed sound-speed field.

For a resolved ray path that connects a source-receiver pair, the time required for a signal to reach the receiver from the source is given by

$$\bar{t}+\delta t = \int_{\underline{x}+\delta\underline{x}} \left(\bar{c} + \delta c + \underline{v}.\frac{d(\underline{x}+\delta\underline{x})}{ds} \right)^{-1} ds, \qquad (3.7)$$

where the quantity in the bracket is often referred as the ray speed, $\bar{t}$ is the travel time in the unperturbed and motionless state

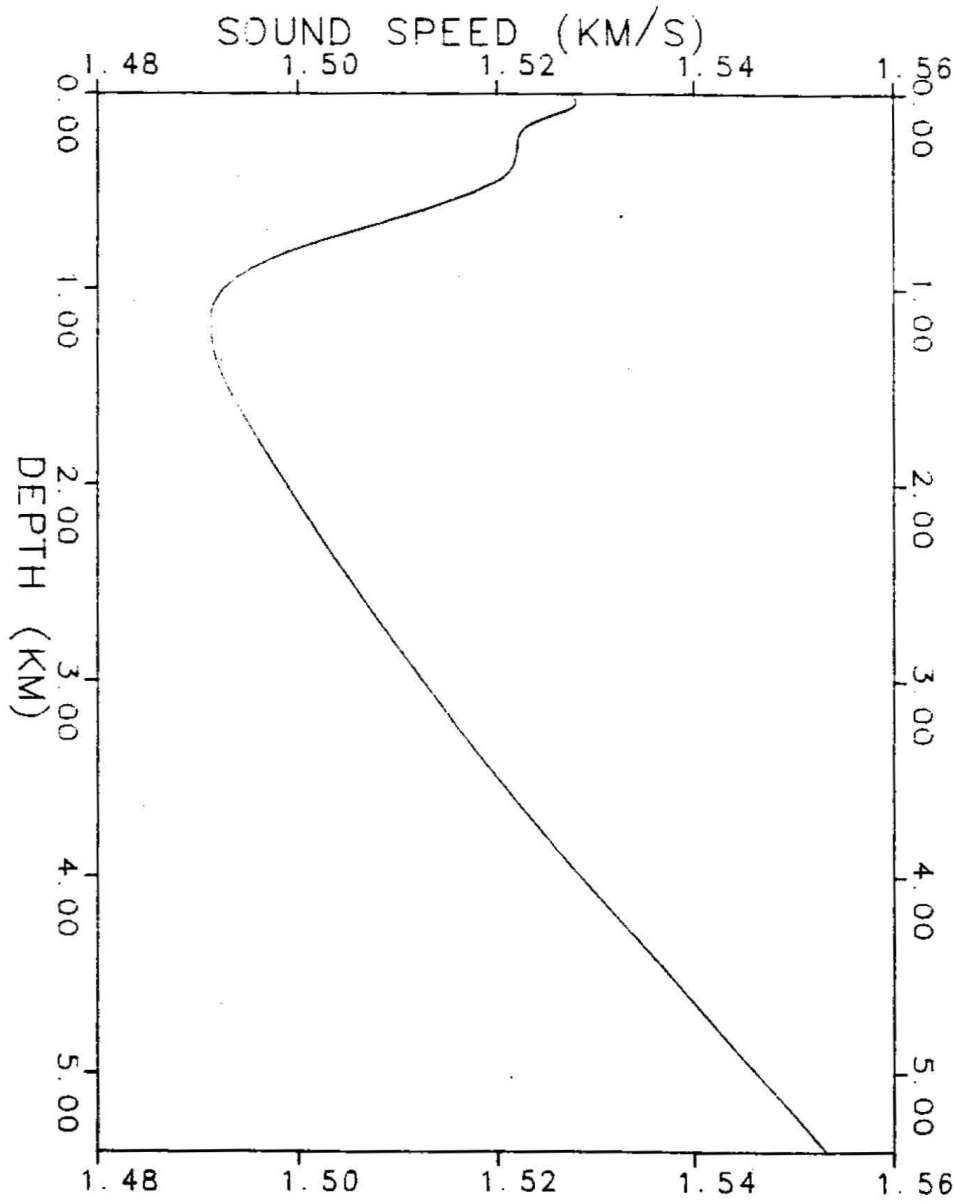

Figure 3.2a. The averaged sound-speed profile in the tomographic region.

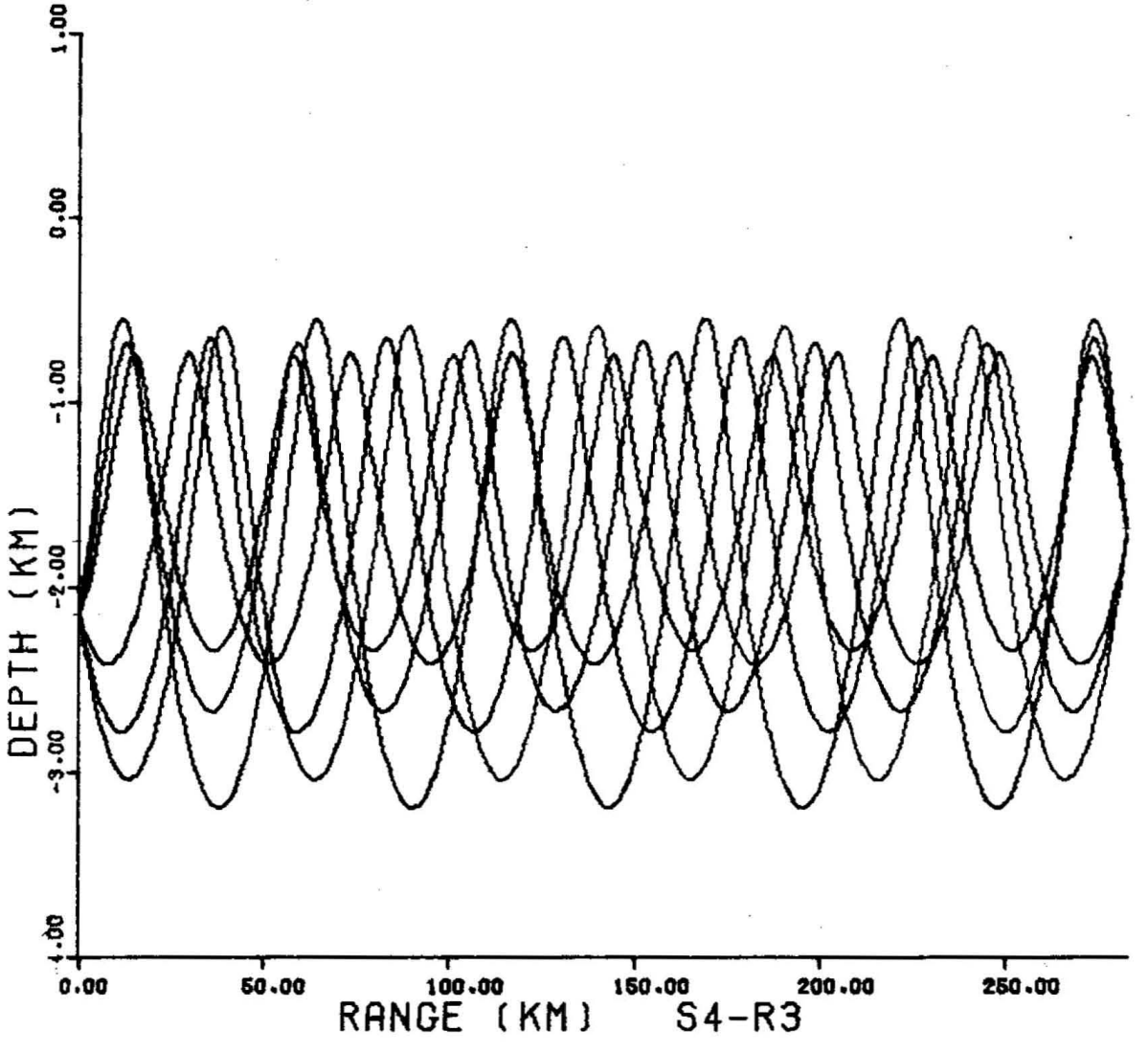

Figure 3.2b. A ray diagram, showing the paths of 3 of the eigen-rays that connected the source S4 to the receiver R3.

and δt is its deviation due to δc and $\underline{v}$. It is seen that in general travel times are perturbed in a very complicated manner. Both the sound-speed perturbations and the currents can affect travel times directly and they can also affect travel times indirectly by changing the trajectories of the ray paths.

The evaluation of δt can be simplified. Hamilton et al. (1980) have shown that for any stable ray, that is any ray which exists in the mean state and does not disappear or alter drastically its geometry in the perturbed state, and for weak horizontal variations in c and $\underline{v}$, the contribution of $\underline{\delta x}$ to δt is of higher order than that contributed explicitly by the changes in the ray speed. Therefore, they concluded that the perturbed travel times may be evaluated along the unperturbed ray paths without losing much accuracy. Furthermore, for most oceanic fluctuations $|\delta c| >> |\underline{v}|$ and hence $\underline{v}$ may be neglected together with $\underline{\delta x}$. By further neglecting terms of order $(\delta c/c)^2$, δt may be approximated by

$$\delta t = \int_{\underline{x}} \frac{-\delta c}{c^2} \, ds. \tag{3.8}$$

In (3.8), δt represents an integral measurement of δc. Because of the averaging process, oceanic fluctuations of smaller scales are automatically filtered from δt. This is one of the many advantages of acoustic techniques over traditional techniques of spot measurements.

Although Mercer and Booker (1983) have found conflicting evidence for the validity of the assumption of travel-time linearity (3.8) for the case of a warm eddy at temperature changes greater than 1°C, the validity of (3.8) for the case of planetary waves and ranges of 300 km were confirmed by us through a computer simulation. Planetary waves that correspond to δc of order 5 m/s and $\underline{v}$ of order 5 cm/s in the upper ocean were simulated; these values are typical of the open ocean. Perturbed and unperturbed ray paths over a distance of 300 km that connects a source point and a receiving point on the channel axis were computed by solving (3.6) numerically with a fourth-order range-dependent ray-tracing technique, developed by the author using the Runge-Kutta method (Acton, 1970) and thus obtaining high numerical accuracy at long range. The travel-time perturbations were then computed numerically from both (3.7) and (3.8), and comparisons made. Results of the simulated study are summarized as follow:

(1) Travel-time perturbations of order 30 ms are found.

(2) Ray paths are practically unperturbed. The vertical and horizontal changes of their geometries are of orders 50 m and 1/2 km, respectively. These changes are small comparing to the scales of the mesoscale perturbations. Furthermore, negligible errors of order 3 ms are introduced in δt when the unperturbed ray paths are used.

(3) Current effects are negligible. Travel-time perturbations created by the flow field are found to be of order 2 ms.

Although the total error created by the assumption of stationary ray paths and the neglect of current effects can be more than 10 percent of the signal, the estimate of a few unknown parameters is generally unaffected by the error when a large number of travel-time data are available.

3.3 Data Used

It is well-known that at lower mid-latitudes the mixed upper layer of the water is well separated from the rest of the ocean by a sharp seasonal thermocline located at a depth of about 200 m (Pickard and Emery, 1982). This large and sudden change in the density profile, that is the seasonal thermocline, may be viewed in Fig. (3.3) in which we show an average profile of the buoyancy frequency N(z) in the tomographic region (N^2 is proportional to the density gradient). Physically, the seasonal thermocline inhibits significant exchange of energy between the mixed layer and the lower ocean that includes the main thermocline zone (approximately from a depth of 300 m to a depth of 1500 m) and the deep zone (below 1500 m depth), so although the upper layer is strongly forced by the atmospheric disturbances, the lower ocean can be left unforced. Thus, an idealized unforced ocean model may be used to describe the dynamics of planetary waves in the entire ocean column except the upper layer.

For these reasons and because the potential energy of the waves is well contained in the main thermocline zone, we did not use time records of temperature and travel time that contain information on the forced fluctuations in the upper layer or the unenergetic signals from the deep zone. That is, the time series records of δc that were obseved in the upper layer or the deep zone and the resolved ray paths that cycled into the upper layer were not used.

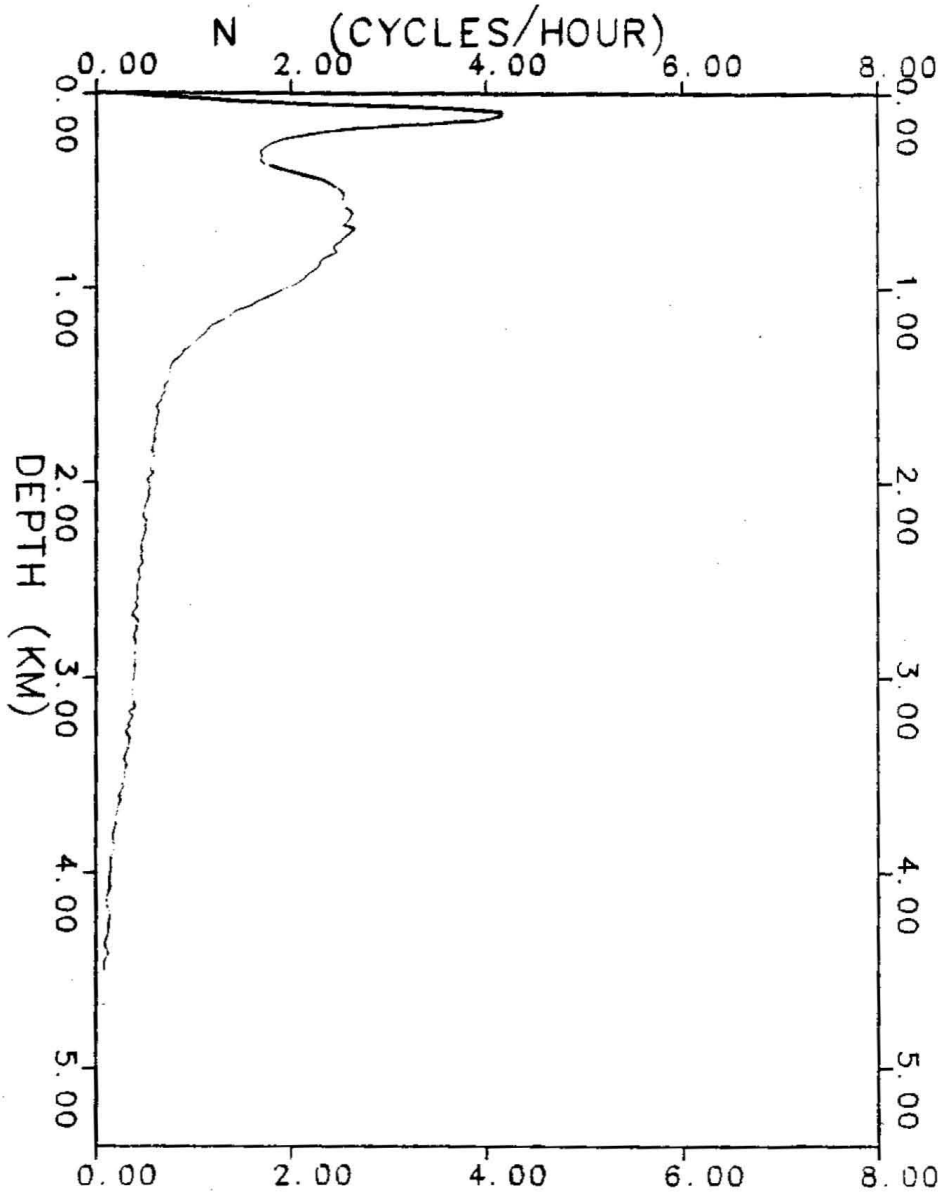

Figure 3.3. The averaged profile of the Brunt-Vaisala frequency in the tomographic region.

(Note that the moored temperature records have been converted to δc time series.) Consequently, we have eliminated all but 7 of the δc time series and 58 of the δt time series in the estimations. We have also checked the deep δc time series (below 1600 m): they have very little variance, which is consistent with the theory.

Some statistical information about the mesoscale variability in the general area of the tomographic region was available from previous experiments, in particular, from MODE. Such information concerning time scales and vertical structures can be very helpful in the data processing (such as filtering) needed for reducing the noise level in the data and the size of the data set. Once noise and data are adequately reduced, more accurate and efficient estimates can be obtained. Note that statistical information can also be used to provide additional constraints on the solution of the inverse problem; the accuracy of the estimate is generally improved by their application (see Ch. 4 for discussions).

Daily averaging corresponding to low-pass filtering was performed on the δc and δt time series so that the noise produced by "uninteresting" events such as tides and internal waves is reduced. Furthermore, data points on every third day and on every ninth day in the filtered time series of δc and δt respectively were retained for the estimates. We have not lost any useful information by this reduction of the data because the time scale of the mesoscale motion in the area is known to be of order of 100 days (Richman et al., 1977).

McWilliams and Flierl (1975) have shown that over 90 percent of the kinetic energy in MODE was contained in two empirical orthogonal vertical modes that closely resemble the barotropic and the first baroclinic modes of Rossby waves. Furthermore, Richman et al. (1977) have shown that about 90 percent of the potential energy, again in MODE, was contained in the first three baroclinic modes, with 65 percent of the energy being contained in the first Mode alone. Thus, it is evident that the vertical structure in the region is predominantly composed of only a few of the lower modes. In Fig. 3.4, we show the first three baroclinic modes of currents ($f_i(z)$; i=1,2,3), evaluated numerically from (2.24) using N(z) shown in Fig. 3.3 and normalized according to (2.35); the barotropic-current mode is constant through out the water column and is not shown in the figure. The three corresponding vertical-displacement modes, given by $h_i(z)=Df_0^2N^{-2}df_i/dz$ where D=5.4 km is the nominal depth and $f_0=6.38\times10^{-5}$ s^{-1} is the coriolis parameter of the region, but re-normalized to have maxima of unity, are shown in Fig. 3.5. Because of the dominance of the low modes, the δc profile data can be largely reduced, and the reduction will be discussed next.

The vertical modes of sound-speed perturbation, $g_i(z)$'s, can be evaluated by (2.34b). But due to the fact that the potential sound-speed gradient is proportional to cN^2 (Flatte et al., 1979), (2.34b) can be recast as

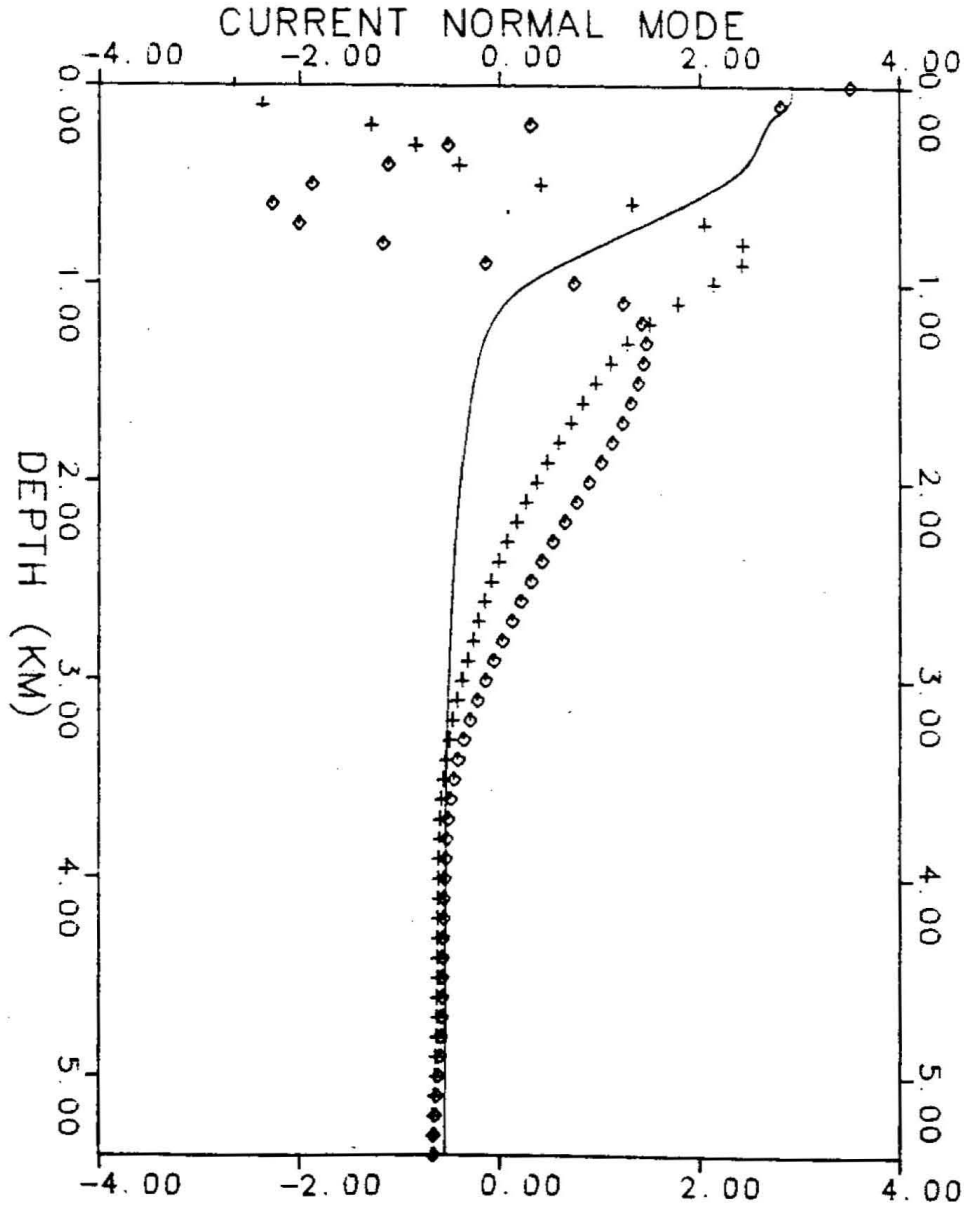

Figure 3.4. The 1st (_____), 2nd (+ + +) and 3rd (◊ ◊ ◊) baroclinic modes of horizontal current in the tomographic region, normalized to have depth-averaged energies of unity.

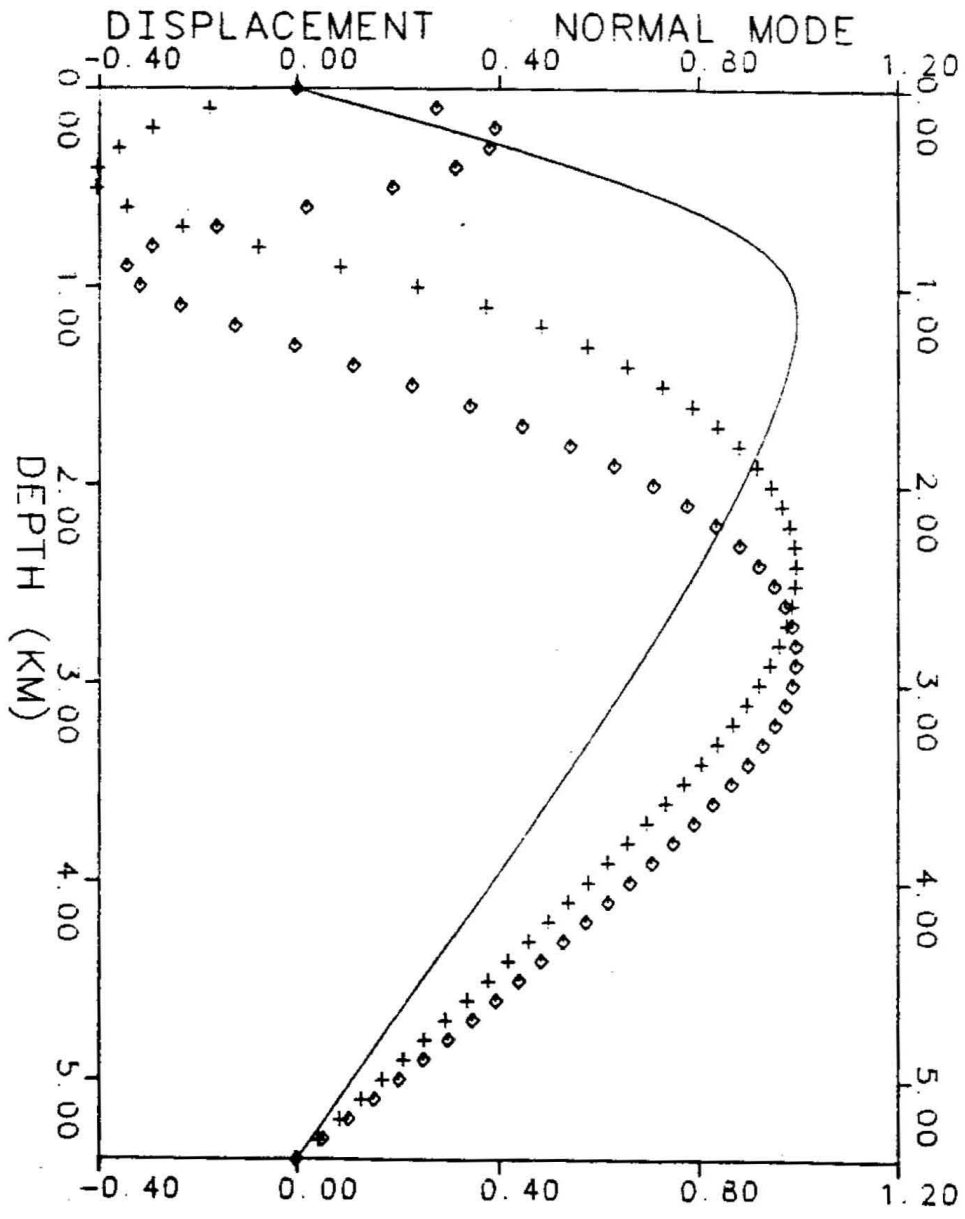

Figure 3.5. The 1st (_____), 2nd (+ + +) and 3rd (◊ ◊ ◊) modes of vertical displacement in the tomographic region, normalized to have maxima of unity.

$$g_i(z) \propto h_i(z)c(z)N(z)^2. \tag{3.9}$$

In Fig. (3.6), the g_i's with i=1,2,3 are shown. Here, we have re-normalized the g_i's to have maxima of unity. Since the g_i's constitute a complete set of functions, the observed profiles of δc can be decomposed as

$$\delta c_j^{CTD}(z) = \sum_{i=1} d_{ij}\, g_i(z);\ j=1,2,3,..., \tag{3.10}$$

where d_{ij} represents the weight of g_i in the jth profile δc_j^{CTD}. It can be computed easily by using the fact that the Nh_i's (or $(cN)^{-1}g_i$'s) are orthogonal to each other. We can interpret d_{ij} as the observed modal amplitude of δc at the location and time $(x,y,t)=(x_j,y_j,t_j)$ of the jth CTD cast. In general, an exact modal representation of δc_j^{CTD} requires an infinite sum in (3.10). However, because of the dominance of the low modes, the sum can be truncated after a few terms without losing any valuable information. In fact, quite to the contrary, the quality of the profile data is improved since the truncation is a filtering process in which the more oscillatory but unenergetic higher modes are totally eliminated. An important consequence of the truncation is that the data of an entire profile can be effectively compressed into a few modal amplitudes that contain

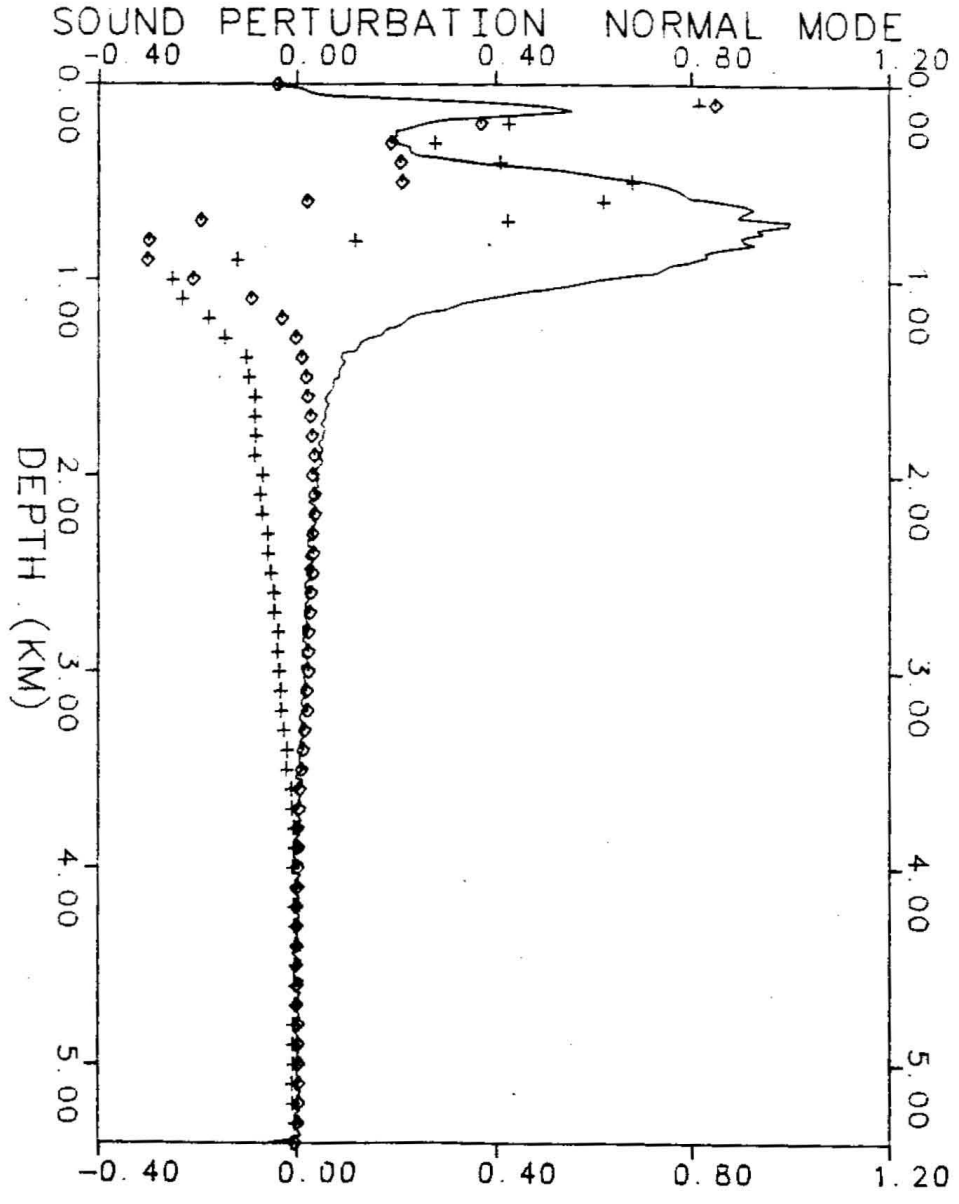

Figure 3.6. The 1st (_____), 2nd (+ + +) and 3rd (◊ ◊ ◊) modes of sound-speed perturbation in the tomographic region, normalized to have maxima of unity.

equivalent information. Therefore, the huge set of δc_j^{CTD}'s can be replaced by a managable set of d_{ij}'s for the lower modes in the parameter estimations.

In order to determine the number of modal amplitudes M required to represent each δc_j^{CTD}, we made the following calculations with M=1,2,3 for each of the casts:

$$P_{Mj} = [\, 1 - \left| \delta c_j^{CTD} - \sum_i^M d_{ij} g_i \right|^2 dz \Big| \Big/ \int \delta c_j^{CTD\,2} dz] \times 100 \text{ percent} \qquad (3.11)$$

where P_{Mj} is the percentage of variance in δc_j^{CTD} generated by the first M baroclinic modes alone. To avoid being misled by the fluctuations in the upper layer, the integrations in (3.11) were performed from 300 m down. Not unexpectedly, P_{1j}'s of 50 to 90 percent were found in all the casts. This finding is consistent with the result of Richman et al. (1977) in MODE. We have also found that the contributions of the 2nd and 3rd modes to δc in the tomographic region are minimal: there being less than a 5 percent increase in the P_{2j}'s and P_{3j}'s from the P_{1j}'s. As a result of the above findings, we have retained only the modal amplitudes of the first mode, that is $a_j^o = d_{1j}$, for the parameter estimates.

It is an interesting fact that even if higher modes do exist and contain significant energy, they are quite transparent to the travel-time measurements. Higher modes are more oscillatory over the vertical column, so that sound waves accumulate many canceling

changes in their travel times as they propagate up and down the ocean along the multipaths before reaching the receiver. To demonstrate this fact, we simulated three perturbed oceans that have the same horizontal scale (of order 100 km) in their δc. The 1st, 2nd and 3rd oceans were perturbed solely by the 1st, 2nd and 3rd modes, respectively. Using the geometry of the 1981 tomographic array and the same 58 ray paths used in the estimations, we computed the corresponding δt's. The rms values of the simulated δc and computed δt's for each ocean are summarized in Table 3.1. It is seen that even with unrealistically large higher-mode perturbations, the second mode is already transparent to the travel-time measurements at an experimental noise level of 5 ms.

Table 3.1

A summary of a simulated study of whether higher modes are transparent to travel-time measurements.

ocean no.	mode simulated	rms δc (m/s)	rms δt (ms)
1	1st	2.1	28
2	2nd	.71	4.2
3	3rd	.52	1.8

We now summarize the data set used in the parameter estimates in Table 3.2. The seven time series records of δc were distributed only at 3 mooring sites E1, E2 and S3, and are thus expected to mainly contain information on the time behaviour of the perturbed field. In contrast, since the duration of each CTD survey is relatively short (2.5 weeks) as compared to the wave period (of order 100 days), they should mainly contain spatial information. About three ray paths per source-receiver pair (which cycle almost the entire depth of the main thermocline zone) were used. The corresponding time series records of travel time therefore contain information on both the time and space behaviour of the perturbed field. Only the data obtained within the period between yeardays 61 and 139 are used since most of the acoustic instruments had failed after yearday 139 and the experiment started roughly on yearday 61. Thus, the data set contains information on the mesoscale perturbations that is continuous in both time and space in the 300 km square over a period of 80 days. The postion 26°N, 70°W and the time yearday 66 are defined hereafter as the point (x,y,t)=(150 km,150 km, 0 s) in the tomographic experimental coordinate system.

Table (3.2)

Data used in the parameter estimations

data type	notation	quantity	duration (yeardays)	no. of data	source
modal amplitudes	a_j^o	65	66-83	65	1st CTD survey
modal amplitudes	a_j^o	65	120-137	65	2nd CTD survey
δc time series	δc_{jk}^o	7	61-139	7x27	temperature sensors
δt time series	δt_{jk}^o	58	61-133	58x9	tomographic array

Note that j is the index for position and k is the index for time.

3.4 The Wave-Induced Sound-Speed Perturbations

Propagating planetary waves can be affected by a number of factors, such as the presence or absence of a mean flow, a bottom slope or resonant interactions. Depending on which of those effects are important, the corresponding perturbed field can display very different space-time characteristics. Due to the uncertainty on which of those effects dominate in the real situation, different but plausible dynamical models that place emphasis on different factors and are parameterized by different sets of wave parameters must be tried in the detection process. Thus the detection of planetary wave involves both parameter estimation and model identification.

For the detection of baroclinic waves in the tomographic region, we have estimated the wave and mean-flow induced sound-speed perturbations $\delta c_m(\underline{x},t;\underline{p})$ both from our three plausible wave-propagation models (labeled 0, 1 and 2) and the data set. The results of the wave-parameter estimation and the goodness of each model are presented and discussed in Ch. 5. In this section, we describe the three models, their associated δc_m's and the corresponding sets of wave parameters $\underline{p}$.

The ocean bottom in the area of the experiment is quite flat so that minimal topographic effects on the wave dynamics should be expected. Thus, in all three models, the modification of the β-effect resulting from depth variations is excluded. The forced waves resulting from nonlinear interaction of the dispervsive waves

are also excluded in all the models. The forced perturbations are of higher order so that they should be negligible. However, at resonance, the forced perturbations can grow in time and hence can become significant, thus the possibility of resonant propagation is included in Model 2. Only the 1st baroclinic waves are modeled because little energy in the higher modes are found in the CTD casts. Furthermore, the waves are assumed to be narrow band so that locally we can use a discrete-wave representation.

Model 0 represents free propagation of linear Rossby waves over a flat bottom in the absence of a mean flow. The isopycnal surfaces are displaced by the baroclinic waves so that the corresponding sound-speed perturbations are given by

$$\delta c_m(\underline{x},t;\underline{p}=\underline{p}_w) = \delta c_w(\underline{x},t;\underline{p}_w) \tag{3.12}$$

with

$$\delta c_w = g_1(z) \sum_{i=1}^{W} A_i \cos(k_i x + l_i y - \sigma_i t + \gamma_i), \tag{3.13}$$

where W is the number of first baroclinic waves considered and A_i, (k_i,l_i), σ_i and γ_i are the amplitude, wavenumber vector, frequency and phase of the ith wave, respectively. The wave amplitude A_i represents the maximum δc (which occurs at z=-700m)

induced by the ith wave since $g_1(z)$ has been re-normalized to have a maximum of unity that occurs at 700 m depth. The space-time behavior of δc_w is characterized by the wave parameters $\underline{p}=\underline{p}_w$ and constrained by the modal dispersion relationship of the waves as given in (2.49) without β modifications and Doppler shifts, i.e., $\delta\beta_n=\delta\sigma_n=0$. Because σ_i is constrained by (k_i,l_i) in (2.49), σ_i is not a free parameter, so that δc_w is completely determinable and can be parameterized by

$$\underline{p}_w = (A_1,k_1,l_1,\gamma_1,\ldots,A_W,k_W,l_W,\gamma_W). \tag{3.14}$$

The possibility of the existence of a mean flow is added in Model 1. The structure of the mean current is assumed to consist of the barotropic and the first baroclinic modes only. This assumption is probably a good one because the two modes are known to contain the greatest fraction of the kinetic energy in this general area (McWilliams and Flierl, 1975, and Sanford, 1975). In this model the isopycnal surfaces are further tilted by the baroclinic mean current (the thermal wind relation). Therefore, the corresponding sound-speed perturbations are now represented by

$$\delta c_m=\delta c_w(\underline{x},t;\underline{p}_w,u_0,v_0,u_1,v_1)+\delta c_c(\underline{x};u_1,v_1,b_0) \tag{3.15}$$

with an additional time-independent mean variation

$$\delta c_c = g_1(z)\left[b_0 - \frac{u_1}{F}\left(\frac{y}{D}\right) + \frac{v_1}{F}\left(\frac{x}{D}\right)\right], \tag{3.16}$$

where F is a known constant, b_0 is a constant for the shifting of the zero-reference of δc_c from the origin $(x,y)=(0,0)$ to the correct position, and (u_0,v_0) and (u_1,v_1) are the modal amplitudes of the barotropic and baroclinic mean currents, respectively. (Note that the overbars on the mean modal amplitudes have been dropped and F is an adjusting factor resulting from the different normalizations of f_i's and g_i's; F=0.157 in the tomographic region.) Due to the Doppler effects, the dispersion relationship of the waves changes from that of Model 0; therefore, so does the space-time behavior of δc_w. The Doppler shifts $\delta\sigma_n$ in (2.49) now exist and are constrained by the (k_i,l_i)'s, (u_0,v_0)'s and (u_1,v_1)'s as given in (2.48C). Thus, δc_m is now parameterized by $\underline{p}=(\underline{p}_w,u_0,v_0,u_1,v_1,b_0)$.

In Model 2, the possibility of the propagation of resonant secondary waves is further included. The modeling requires the replacement of A_i by A_i+G_it in (3.13) where G_i represents the growth rate of the ith wave. In general, G_i is constrained by the wavenumber vectors and wave amplitudes of the interacting primary waves. However, since the barotropic mode is not observable in our data set while resonant waves can be generated by intermodal

wave-wave interactions, G_i can only be left as a free parameter in the model. The set of parameters are now given by

$\underline{p}=(\underline{p}_w, G_1, \ldots, G_W, u_0, v_0, u_1, v_1, b_0)$.

The dynamical assumptions made in each model are summarized in Table 3.3. In addition to the correct propagation model and its parameter values, the number of existing waves W is also an unknown. Therefore, its integer value must also be estimated in the process of detection. The estimation of W is achieved through assumption and parameter estimation, followed by model identification, with each presumed value of W being considered as giving a different sub-model.

Table 3.3

The Dynamical Hypothesis Made In Each Wave-Propagation Model

Model no.	weak mean flow		wave-wave	resonant	flat
	oth mode	ist mode	interactions	propagations	bottom
0					X
1	X	X			X
2	X	X		X	X

"X" denotes the assumption is made.

3.5 The Model Equations

Detection is the extraction of the desired signal from a background of noise (or other signals) by utilizing estimation methods. In our case the desired signals are the sound-speed perturbations δc induced by the waves and the mean-flow.

Obviously, not all the perturbations of sound speed are caused by the baroclinic waves and mean currents. There are many other oceanic events such as tides, internal waves, turbulence, etc. that also perturb the sound speed. These other sound-speed fluctuations, therefore, constitute the background noise of our detection problem, and just like the measurement noise, they too contaminate the data set. But if the signal generated by the planetary waves and mean flow is dominant in the data, the signal can be detected.

The contamination in the data set caused by the background sound-speed fluctuations is referred to as the model noise. The model and measurement noise combine to give the experimental noise that accounts for all the noise in the model equations for the modal-amplitude data and the δc time records. For the qth modal-amplitude datum $a^o_q = d_{1q}$ observed at $(x,y,t)=(x_q,y_q,t_q)$ and the kth datum $\delta c^o_{1k}=\delta c^o_1(t=t_k)$ which is the 1th δc time record observed at $(x,y,z,t)=(x_1,y_1,z_1,t_k)$, the corresponding model equations can be expressed simply as

$$a_q^o = a_q^m(\underline{p}) + v_q^a \tag{3.17}$$

and

$$\delta c_{1k}^o = \delta c_{1k}^m(\underline{p}) + v_{1k}^c, \tag{3.18}$$

where

$$a_q^m = \delta c_m(x_q, y_q, z=-700\ \text{m}, t_q) \tag{3.19}$$

and

$$\delta c_{1k}^m = \delta c_m(x_1, y_1, z_1, t_1) \tag{3.20}$$

are the signals, and v_q^a and v_{1k}^c are the noise in a_q^o and δc_{1k}^o, respectively.

The formulation of the model equations for the δt time series requires some special care. The content of the δt data is more complicated than that of the other data. In addition to the baroclinic waves and mean current, and the background oceanic fluctuations and the measurement errors, the relative motions and the uncertainty in the nominal positions of the acoustic moorings also contribute to the observed travel-time perturbations. In fact, the latter two contributions were dominant. If one were to model these mooring-position related travel-time perturbations as part of the experimental noise, the δt time records would suffer a vanishingly small ratio of signal to noise. In order to improve the

quality of the δt data, as suggested by Cornuelle (1983), the mooring-position related travel-time perturbations must also be modeled as signals, implying that the uncertainty in the mooring positions must also be parameterized in the acoustic model equations.

A set of relative mooring-displacement data was available from the acoustic navigation systems. The tracking data had already been used to eliminate some of the signal produced by the mooring motions in the travel-time data. But, since the set of tracking data is neither error-free nor complete (a lot of data were missing), the untracked or unknown horizontal displacements together with the uncertainty in the horizontal nominal positions of the moorings must still be parameterized. Note that the vertical translations of the acoustic sources and receivers were small (of order 50 m) and produced very little travel-time perturbations (of order 1 ms), therefore, they need not be parameterized.

Let us consider the jth ray path connecting the mth source Sm to the nth receiver Rn. According to Cornuelle (1983), the additional time required for the acoustic wave front to travel from Sm to Rn along the path due to a small elongation δR (let's say of order 1 km) of the horizontal distance separating Sm and Rn can be expressed, to lowest order, as

$$\delta t_j^R = r_j \delta R, \tag{3.21}$$

where r_j is the corresponding ray parameter, i.e., the cosine of

the launching (receiving) angle divided by the sound speed at the source (receiver); r_j is a conserved quantity along the ray. Let the unknown horizontal-displacement vectors at time t_k and the time-independent errors on the assumed nominal horizontal-postion vectors of Sm and Rn be $[\delta x_{Sm}(t_k), \delta y_{Sm}(t_k)]$ and $[\delta x_{Rn}(t_k), \delta y_{Rn}(t_k)]$, and $(\Delta x_{Sm}, \Delta y_{Sm})$ and $(\Delta x_{Rn}, \Delta y_{Rn})$, respectively. It then follows that the corresponding δt_j^R at time t_k is given by

$$\delta t_j^R(t_k) = r_j \cos\phi_{mn} [\Delta x_{Rn} - \Delta x_{Sm} + \delta x_{Rn}(t_k) - \delta x_{Sm}(t_k)] + r_j \sin\phi_{mn} [\Delta y_{Rn} - \Delta y_{Sm} + \delta y_{Rn}(t_k) - \delta y_{Rn}(t_k)] \tag{3.22}$$

where ϕ_{mn} is the direction of the horizontal line of transmission from Sm to Rn, measured in degrees (positive anticlockwise) with respect to the x-axis, i.e., east-axis.

We are now in a position to write down the acoustic model equations. For the travel-time perturbation $\delta t_{jk}^o = \delta t^o(t=t_k)$ observed from the jth ray path at time t_k, the corresponding equation can be cast symbolically as

$$\delta t_{jk}^o = \delta t_{jk}^m(\underline{p}, \Delta x_{Sm}, \Delta y_{Sm}, \Delta x_{Rn}, \Delta y_{Rn}, \delta x_{Sm}, \delta y_{Sm}, \delta x_{Rn}, \delta y_{Rn}) + v_{jk}^t \tag{3.23}$$

where v_{jk}^t represents the total or the experimental noise in δt_{jk}^o.

The signal δt^m_{jk} can be written as the sum of two parts such that

$$\delta t^m_{jk} = \delta t^p_{jk} + \delta t^R_{jk} \tag{3.24}$$

where $\delta t^R_{jk} = \delta t^R_j(t_k)$ is expressed in (3.22) and δt^p_{jk} is the signal induced by the waves and mean flow which can be expressed as, using (3.8),

$$\delta t^p_{jk} = \int_{x_j(s)} \frac{-\delta c_m(\underline{x}, t_k; p)}{\bar{c}(z)^2}\, ds, \tag{3.25}$$

with $\underline{x}_j$ denoting the unperturbed trajectory of the jth ray path.

CHAPTER 4

PARAMETER ESTIMATION AND THE GENERAL NONLINEAR PROBLEM

For a typical scientific investigation, parameter estimation (or inversion) and model discrimination are the two crucial steps in the process of extracting information from data obtained in experiments. Of course, a successful investigation also depends critically on the understanding of the physical situation and the planning of the experiments. While the physical knowledge enables us to develop plausible mathematical models, relating the physical parameters that characterize the physical situation to the pre-selected types of observations (the forward problem), well designed experiments provide good data which are informative to the investigation. Readers interested in the design of experiments are refered to the works of Box et al. (1959, 1963 and 1967).

Estimation theory plays a vital role in making progress in physical oceanography. The ocean is a very complicated environment. The forcing, initial conditions, and boundary conditions are uncertain. The exact description of the fluid motion by mathematical equations is often very difficult, and even when where it is possible, the exact solution is often intractable. Thus, in the theoretical study of an oceanic phenomenon, we must resort to assumptions and approximations (idealizations) that are reasonable for the particular study. Different assumptions and approximations result in different models, and only after

experiments are conducted and estimations performed, can we then compare models for the confirmation, rejection or revision of hypotheses. Therefore, estimation, which utilizes data observed in experiments, provides a feed back loop in the process of understanding the ocean.

This chapter considers the general estimation problem. The technique and results of estimation specific to the observations of planetary waves in this study are presented in chapter 5. The corresponding forward problem has been studied in chapters 2 and 3. In the first part of this chapter, estimation methods developed from pure stochastic approaches as well as those with few probabilistic considerations are reviewed and discussed. Our goal is to relate and unify these methods by showing that once the same set of information and assumptions concerning the solution and experimental noise is consistently and analogously adopted by each individual method, these methods give the same solution. In showing this, a generalized estimation procedure that computes this "optimal" solution common to all the methods considered is also established. The implication is that we can stop worrying about these different methods and just apply the generalized procedure to data, since the solution is independent of the methods themselves. The generalized procedure is the minimization of the now familiar function of a weighted sum of products of residuals from both the experimental and "a priori" data. The second part of this chapter reviews and discusses some widely used minimization methods for computing the

solution. The last part considers the errors in the solutions and presents some overall measures of goodness of a model based on its final residuals. Such measures of goodness are needed in comparing models.

4.1 The General Estimation Problem

The models for many physical situations can be expressed symbolically as

$$\underline{y} = \underline{f}(\underline{x},\underline{p}), \tag{4.1}$$

where $\underline{y}$ is an observable vector representing the signal produced by a physical event, $\underline{x}$ is a controllable vector of design-parameters defining the experimental conditions for observing $\underline{y}$, $\underline{p}$ is a vector of physical parameters, where the value of $\underline{p}$ is not of our choosing but rather characterizes the physical event, and $\underline{f}$ is a vector of functions (model equations) which express one's theory on the relation among quantities; $\underline{f}$ is a vector of functionals when $\underline{p}$ represents continuous functions in their parametric forms. Let us define the dimensions of $\underline{y}$, $\underline{f}$, $\underline{x}$ and $\underline{p}$ to be mx1, mx1, rx1 and nx1, respectively. Note that all the vectors are column vectors.

The study of a forward problem, typically, consists of identifying a relevant set of physical parameters and deriving an appropiate set of model equations. The idea is to be able to model the signal for a given situation described by $\underline{x}$ and $\underline{p}$, by incorporating all the essential features of the true process into $\underline{f}$.

The corresponding inverse problem is the estimation of $\underline{p}$, based on data obtained in an experiment of controlled $\underline{x}$. The model equations $\underline{f}$ are considered to be known from the study of the forward

problem. In the presence of additive random noise in the observations, which is always the case in practice, the experiment can be modeled stochastically as

$$\underline{y}^* = \underline{f}(\underline{x},\underline{p}) + \underline{v}. \tag{4.2}$$

The data or observations, denoted as $\underline{y}^*$, contain the signal, but unfortunately, are contaminated by noise $\underline{v}$, and $\underline{y}^*$ and $\underline{v}$ are both (m-dimensional) vector random variables. The experimental noise $\underline{v}$ includes both the measurement noise and model error. Because the data is imperfect, only approximate solutions or estimates are obtainable. An estimation or inversion procedure acting on the data to give an estimate is called an estimator. In general, different estimates may or may not be computed from different estimators, given the same data set. However, the "optimal" or the "best" estimate $\underline{p}^*$, that is the unique solution for $\underline{p}$, is evaluated from the optimal estimator which is established according to one's criteria for the optimal estimate. Consequently, the quality of $\underline{p}^*$, besides depending on the quality of the observations and the model, depends on the estimator that is employed.

4.2 Establishing Stochastic Estimators

Due to the randomness in $\underline{y}^*$, although $\underline{p}$ itself may or may not be a vector random variable (a random p corresponds to a random process), the estimates are always random in nature: For a given estimator, different realizations of $\underline{y}^*$, or equivalently, of $\underline{v}$, would result in different estimates. In fact, the statistical properties of the estimates depend on the estimator used and the statistical properties of $\underline{v}$.

Before establishing the estimator for computing the optimal estimate $\underline{p}^*$, one must do the following: (1) Select a desired set of statistical criteria for the optimal estimate, (2) collect all the available statistical information concerning the noise $\underline{v}$, and (3) collect all the prior information concerning the physical parameters $\underline{p}$.

4.2.1 Criteria For The Optimal Estimate

A reasonable estimator should produce estimates which, on the average, are close to the true value of $\underline{p}$. There are two types of error associated with $\underline{p}^*$: the bias and the random errors, and small bias and small variance are generally highly desirable. (Bias is the difference between the expected value of the estimate and the true solution.)

In most cases, unbiased estimators are hard to obtain, and even

if obtainable, the corresponding estimates are usually unstable to noise, meaning that small errors in the data can be translated into large errors in the estimate. In fact, a small bias must often be introduced, intentionally, for uniqueness and for reducing the variance of the solution of an ill-conditioned system (Rust and Burrus, 1972) (here "system" means system of equations as expressed in equation (4.2)). Thus, total lack of bias is neither essential nor often desirable, because unbiased estimates are not error-free and are sometimes unstable.

The theoretically attainable lower bound of variance is given by the Rao-Cramer theorem (see Bard, 1974, for the derivation). However, practically, the estimator associated with this minimum variance bound (MVB) can only be established for a few simple systems such as linear systems. The MVB estimators in the case of linear systems can be derived easily by the Gauss-Markov theorem (Liebelt, 1967). In many engineering applications of estimation theory, the development of a new estimation method is usually not necessary or important, because many of the existing and commonly used estimators can generally provide reasonably accurate estimates, and in addition, the minimum-variance, unbiased (i.e. the most ideal) estimator is unattainable in most cases, anyway. In choosing a common method, we have simply accepted the criteria for the optimal estimate associated with the method.

4.2.2 Noise Distribution

Complete statistical knowledge of the vector random variable $\underline{v}$, that is its joint probability distribution function (pdf) is seldom possessed, and usually only a little information concerning v is available, for example, its mean (vector) and covariance (matrix). However, we must somehow assign to $\underline{v}$ an adequate pdf because most estimation methods demand it. If only the mean and variance are known, a rational choice is the (multivariate) Gaussian (or normal) distribution, for reasons stated in the following paragraphs:

(1) Simplicity: a Gaussian distribution is parameterized by its mean and variance only, and the assumption of a normal pdf for $\underline{v}$ generally leads to the establishment of simple estimation procedures.

(2) Under some mild conditions, if $\underline{v}$ is generated from a summation or integration of many random variables, whether normal or not, $\underline{v}$ tends to the normal, according to the Central Limit theorem (a proof of the theorem can be found in Drake, 1967).

(4) We do not want the estimator to be falsely informed by specifying more statistical information than we actually know. In information theory, Shannon (1948) has derived a suitable measure of the information contained in a pdf; this measure is called the entropy and it is inversely proportional to the amount of information. Without further information beyond the mean and variance, the Gaussian distribution maximizes the entropy (the proof

can be found in Bard, 1974), implying that the amount of extraneous information is minimized.

Henceforth, $\underline{v}$ is assumed to be normally distributed with zero mean and a known nonsingular symmetric covariance matrix $\underline{C}_{\underline{v}}$. In many cases, the true $\underline{C}_{\underline{v}}$ may not be exactly known, but this poses no serious problem in the estimate. In general, a reasonalbe approximation of $\underline{C}_{\underline{v}}$ can suffice, because most estimators are not sensitive to small variation in $\underline{C}_{\underline{v}}$. In addition, the parameters can always be reestimated using a refined $\underline{C}_{\underline{v}}$ when the noise estimate (i.e. the final residuals) generated by the estimator signifies that the original specification is far from being correct. Model errors, very often, have nonzero means, and the assumption that the means are zero will result in the generation of bias error in the estimate. However, when the model is accurate, the bias will be small. The generation of bias will be further discussed in Ch. 6, Sec. 6.4.

4.2.3 Prior Information

Prior information, if available, can often increase the accuracy of the estimate. In fact, for ill-conditioned systems, the use of prior information, which is equivalent to the introduction of bias in the case of linear systems, must be insisted upon (Jackon, 1979, Rust and Burrus, 1972; also see the discussion in Ch. 6, Sec. 6.4

for reasons of stability and uniqueness. The information can come from previous experiments or physical intuition, and it can generally be summarized in two forms: a priori probability distributions, and deterministic equality and inequality contraints for $\underline{p}$.

A scientist usually has some idea of the true value of p before carrying out an experiment. For instance, he may know that the true $\underline{p}$ must lie in a region around, say, $\underline{p}=\underline{p}_0$. The above information can often be expressed by inequality contraints. On the other hand, if one is willing, the same information can be summarized in an a priori pdf $P(\underline{p})$: The specification of the a priori expectation by $\underline{p}_0$ and the a priori covariance matrix $\underline{C}_{\underline{p}}$ according to the boundary of the region leads naturally to the assignment of an a priori Gaussian distribution for $\underline{p}$, with respect to information theory. In what follows, we consider only estimation with a priori probability distribution.

4.3 Statistical Estimation Methods

Maximum likilehood (ML) and the mode of the posterior distribution (MPD) (when statistical prior information is available) are representative, common estimation methods. One important reason for their typicality is that many other common methods give the same estimate when all the random variables under consideration are normal. Other reasons are their wide range of utility, simplicity in applications and that the estimates are generally easy to compute.

The ML estimate (MLE) is the value of $\underline{p}$ that maximizes the likelihood function obtained by substituting the realization of $\underline{y}^*$ into $P(\underline{y}^* | \underline{p})$, i.e., the pdf of $\underline{y}^*$, given $\underline{p}$. The reasoning is that the MLE is associated with the physical event which is most likely to produce the data that we have observed. On the other hand, the MPD estimate (MPDE), as indicated by its name, is just the value of $\underline{p}$ at which the maximum of the a posteriori distribution $P(\underline{p} | y^*)$ occurs; $P(\underline{p} | \underline{y}^*)$ is the pdf which we must assign to $\underline{p}$ after the experiment was conducted, that is the pdf of $\underline{p}$, given $\underline{y}^*$. Clearly, the MPD method is simply an extension of the idea of maximum likelihood to accommodate the use of prior information. Both the MLE and MPDE are asymptotically unbiased (or consistent) and asymptotically efficient (Fisher, 1950), that is the estimates become unbiased and reach the MVB when the number of observations increases to infinity, therefore, we would expect the estimates to

have small biases and small variances when the set of data is much larger than the set of parameters. Futhermore, the estimates do not depend too strongly on the actual shapes of the distribution functions $P(\underline{y}^* | \underline{p})$ or $P(\underline{p} | \underline{y}^*)$, and the tails of the distributions have no effect at all on the estimates.

The MPD method belongs to the class of estimation methods which uses Bayes' theorem. Let us now formulate the MPD estimator. From Bayes' theorem, we have that

$$P(\underline{p} | \underline{y}^*) = P(\underline{y}^* | \underline{p}) P(\underline{p}) / P(\underline{y}^*). \tag{4.3}$$

But since

$$P(\underline{y}^*) = \int P(\underline{y}^* | \underline{p}) d\underline{p} \tag{4.4}$$

is not a function of p, and together with the assumption of normal distributions such that

$$P(\underline{y}^* | \underline{p}) = (2\pi)^{-m/2} \det{}^{-1/2}(\underline{C}_{\underline{v}}) \, e^{-1/2[\underline{y}^* - \underline{f}(\underline{x},\underline{p})]^T \underline{C}_{\underline{v}}^{-1} [\underline{y}^* - \underline{f}(\underline{x},\underline{p})]} \tag{4.5}$$

and

$$P(\underline{p}) = (2\pi)^{-n/2} \det{}^{-1/2}(\underline{C}_{\underline{p}}) \, e^{-1/2(\underline{p}-\underline{p}_0)^T \underline{C}_{\underline{p}}^{-1} (\underline{p}-\underline{p}_0)}, \tag{4.6}$$

it follows that the maximum of $P(\underline{p}|\underline{y}^*)$ is identical to the minimum of the function:

$$s(\underline{p}) = s_d(\underline{p}) + s_p(\underline{p}), \tag{4.7a}$$

where

$$s_d(\underline{p}) = 1/2[\underline{y}^*-\underline{f}(\underline{x},\underline{p})]^T\underline{C}_{\underline{v}}^{-1}[\underline{y}^*-\underline{f}(\underline{x},\underline{p})] \tag{4.7b}$$

and

$$s_p(\underline{p}) = 1/2(\underline{p}-\underline{p}_0)^T\underline{C}_{\underline{p}}^{-1}(\underline{p}-\underline{p}_0). \tag{4.7c}$$

The function $s(\underline{p})$ is called an objective function, which is a measure of the "lack of fit" between the data and model for a given value of $\underline{p}$ (Bard, 1974). We can interpret s_d and s_p as the constraints on $\underline{p}$ provided by the data and prior information, respectively. Thus, the MPD estimator is the minimization of the objective function of equation (4.7), and the location of the (least) minimum is the MPDE. It is a general fact that almost any estimation method can be reduced to the minimization of an objective function, as will be shown below.

A few comments on the minimum point $\underline{p}^*$ of equation (4.7), that is the MPDE, are listed below:

(1) If prior information is not available so that $\underline{C}_{\underline{p}}^{-1}=0$, $\underline{p}^*$ is identical to the MLE. This can be shown by observing that the

minimum of s_d is the maximum of the likelihood function of equation (4.5).

(2) Even when $C_p^{-1} \neq 0$, $\underline{p}^*$ may be interpreted as the MLE: As pointed out by Jackson (1979), the a priori information may be incorporated in the system of equations by treating $\underline{p}_0$ as the (a priori) data for p with covariance matrix $\underline{C}_p$. In this way, there are n more equations in the system and $\underline{p}^*$ is where the maximum of the modified likelihood function occurs, assuming $\underline{v}$ and $\underline{p}$ are not correlated.

(3) If the model equations $\underline{f}$ are linear in $\underline{p}$ (linear system), and $\underline{v}$ and $\underline{p}$ are uncorrelated, $\underline{p}^*$ is identical to the linear minimum variance (Gauss-Markov) estimate (Liebelt, 1967).

(4) If the data are not enough to constrain $\underline{p}$ by themselves, that is the system is underdetermined and/or ill-conditioned such that more than one least minimum exists when minimizing s_d alone, then additional constraints provided by the prior information denoted by the term s_p must be added to impose uniqueness. This is always the case when inverting functions, for $\underline{p}$ is effectively infinite dimensional. On the other hand, if the system is well-conditioned, which may be the case when the data outnumber the physical parameters, then the addition of s_p in the objective function will have little effect on $\underline{p}^*$.

4.3.1 Incorporation Of Different Data Types

We define an independent data set as a subset of the entire set of data, produced by the same physical event, but measured with a different technique, so that the randomness of any one subset is statistically independent of the other subsets.

Suppose $\underline{y}^*$ is a joint vector of k independent data subsets $\underline{y}_i^*$ acquired in the experiment

$$\underline{y}_i^* = \underline{f}_i(\underline{x}_i, \underline{p}) + \underline{v}_i; \quad i=1,2,\ldots,k, \tag{4.8}$$

where $\underline{f}_i$, $\underline{x}_i$ and $\underline{v}_i$ are respectively the vectors of model equations, design parameters and random noise, corresponding to the observation of $\underline{y}^*$. Since $\underline{v}_i$ and $\underline{v}_j$ are uncorrelated, the data constraint s_d of the objective function of equation (4.7) for the optimal estimate decomposes into a sum of sub-constraints such that

$$s_d = 1/2 \sum_{i}^{k} [\underline{y}_i^* - \underline{f}_i(\underline{x}_i, \underline{p})]^T \underline{C}_i^{-1} [\underline{y}_i^* - \underline{f}_i(\underline{x}_i, \underline{p})], \tag{4.9}$$

where $\underline{C}_i$ is the covariance matrix of $\underline{v}_i$. $\underline{C}_i$ has two important functions: (1) to nondimensionalize the ith set of data and equations so that data sets with different physical units can be incorporated together, and (2) to control the relative effect of the ith data set on the estimate upon its reliability.

4.3.2 Treatment of Erroneous Design Parameters

Optimal values of the design parameters $\underline{x}=\underline{x}_0$ are preselected so as to optimize the effectiveness of an experiment that will be performed. However, the introduction of error in the preselected value of $\underline{x}$ can seldomly be avoided during the deployment. For instance, a physical oceanographer may want to deploy a mooring at a preselected location, but the imperfection in navigation renders the preselected position subject to error. If the signal in data produced by the error in $\underline{x}$ is smaller than the noise level, the error may be of no consequence; otherwise, minimizing the objective function of equation (4.7) will produce an erroneous result which can no longer be an optimal estimate of $\underline{p}$.

This problem can be dealt with by treating the preselected value $\underline{x}_0$ of $\underline{x}$ as the observation of the true value of $\underline{x}$, and modifying the system of equations (4.2) to

$$\begin{bmatrix} \underline{y}^* \\ \underline{x}_0 \end{bmatrix} = \begin{bmatrix} \underline{f}(\underline{x},\underline{p}) \\ \underline{x} \end{bmatrix} + \begin{bmatrix} \underline{v} \\ \underline{w} \end{bmatrix}, \tag{4.10}$$

where $\underline{w}$ is the error in $\underline{x}_0$. In this system, the true value of $\underline{x}$ is also treated as a vector variable to be estimated, and there are additional r unknown parameters and r data points. Suppose we have an idea of what the bounds on $\underline{w}$ are, so that we can characterize $\underline{w}$

by a normal distribution with a covariance matrix $\underline{C}_{\underline{w}}$. This leads to the minimization of the following modified objective function:

$$s(x,p) = \frac{1}{2}[\underline{y}^*-\underline{f}(\underline{x},\underline{p})]^T\underline{C}_{\underline{v}}^{-1}[\underline{y}^*-\underline{f}(\underline{x},\underline{p})]+\frac{1}{2}(\underline{x}_0-\underline{x})^T\underline{C}_{\underline{w}}^{-1}(\underline{x}_0-\underline{x}) + \frac{1}{2}(\underline{p}_0-\underline{p})\ \underline{C}_{\underline{p}}^{-1}(\underline{p}_0-\underline{p}) \qquad (4.11)$$

We refer $\underline{x}_0$ and $\underline{p}_0$ as the erroneous design data and the a priori data, respectively. They result in two constraining functions which are similar in form to those given by the experimental data $\underline{y}^*$.

4.4 Non-probabilistic Estimation Methods

There are estimation methods developed originally with little or no consideration of statistics. These methods do not consider optimal statistical criteria for the estimate, but instead, the optimal criteria are selected in a more deterministic manner upon physical intuition or sometimes in an ad hoc manner. However, these methods have their analogues in the pure stochastic framework once adequate probability distributions are attached. For examples, the primitive method of "weighted least squares" for estimating a handful of numbers is related to the ML method, and the recent "variational method" of Provost (1983) and the classical "inverse methods" of Backus and Gilbert (1967, 1968 and 1970), Wiggins (1972), Jackson (1972), Parker (1977) and Wunsch (1978) for estimating continuous functions are related to the MPD method.

4.4.1 The Variational Method

Provost's variational method translates the problem of estimating continuous functions to a problem in the calculus of variations. In the simplest description, the estimation problem becomes the determination of $\underline{p}$ that minimizes a nonnegative "smoothing" functional of the unknown function represented parametrically by $\underline{p}$, and subject to the data constraint

$$[\underline{y}^*-\underline{f}(\underline{p})]^T \underline{W}[\underline{y}^*-\underline{f}(\underline{p})] = q. \qquad (4.12)$$

In the above, q is an expected (or presumed) positive value of a measure of the total misfit between data and model prediction, and $\underline{W}$ is a positive-definite, diagonal weighting matrix for nondimensionalizing and scaling data and equations having different physical units and different order of magnitudes. Scaling factors associated with the degree of reliability on each datum can also be included in $\underline{W}$. Let's assume that the unknown function is a time signal in this discussion. The smoothing functional can be the integral of the square curvatures or square slopes, or some other desired nonnegative measures of smoothness of the time signal, and it can be expressed parametrically as $\underline{p}^T\underline{S}\ \underline{p}$ with the matrix $\underline{S}$ being positive definite. The corresponding objective function(al) to be minimized is $I(\underline{p},\alpha)$ with

$$I(\underline{p},\alpha)+\alpha q = \alpha[\underline{y}^*-\underline{f}(\underline{x},\underline{p})]^T\underline{W}\ [\underline{y}^*-\underline{f}(\underline{x},\underline{p})] + \underline{p}^T\underline{S}\ \underline{p}, \qquad (4.13)$$

where α is the Lagrange multiplier to be found.

In the variational method, a criterion for the optimal estimate is the satisfaction of the data constraint, but since there will be so many solutions satisfying this constraint due to the underdetermined nature of this system, another criterion must be brought in to ensure uniqueness, and it is smoothness. Clearly, the

method chooses among all the solutions of equation (4.12) the smoothest one to be the optimal estimate, with smoothness defined by the selected smoothing functional.

With $\underline{p}_0=0$, the similarity between equations (4.13) and (4.7) is evident, and in fact, they have the same minimum-point (i.e., the two methods have identical solution) when $\underline{S}$ and $\underline{W}$ are set proportional to the inverse covariance matrices $\underline{C}_{\underline{p}}^{-1}$ and $\underline{C}_{\underline{v}}^{-1}$ of $\underline{p}$ and $\underline{v}$, respectively. Under such choice of $\underline{S}$ and $\underline{W}$, if $\underline{p}$ represents the Fourier amplitudes of the time signal to be estimated, then the diagonal elements of $\underline{S}^{-1}$ and $\underline{W}^{-1}$ represent the normalized power spectral density functions of the signal and noise, respectively. Furthermore, α is analogous to the signal to noise ratio and the deterministic criterion of smoothness is analogous to the a priori information of a low-pass signal described statistically by the spectrum denoted by $\underline{S}^{-1}$.

In practice, one does not compute α and $\underline{p}$ simultaneously through minimization, but instead, they are often determined by an iterative technique: A guess value for α is used so that $\underline{p}$ are the only variables during minimization, and after the corresponding solution for $\underline{p}$ is evaluated, one then computes the corresponding q from equation (4.12), and if the computed q is acceptablely close to the expected value, the optimal estimate is successfully found, otherwise, the procedure is repeated as many times as needed with different but progressively better guess values for α. A similar iterative estimation procedure is also commonly exercised in

stochastic methods because $\underline{C}_{\underline{v}}$ and $\underline{C}_{\underline{p}}$ are generally estimates themselves, therefore, their values must be adjusted and the minimization procedure must be repeated if the final residuals do not agree with the presumed covariances.

4.4.2 The Inverse Methods

Backus and Gilbert originally developed a general formalism for solving the linear inverse problem, which was later cast into simple linear algebra by Wiggins, Jackson, Parker and Wunsch in applications to geophysical and oceanographic problems. Such formalism is, by now, known simply as "linear inverse methods". The general linear inverse problem can be cast into the parametric form

$$\underline{y}^* = \underline{F}\,\underline{p} + \underline{v} \qquad (4.14)$$

by replacing $\underline{f}(\underline{p})$ with $\underline{F}\,\underline{p}$ in equation (4.2), where $\underline{F}$ is a mxn matrix representation of the linear differential operator associated with the forward model and $\underline{p}$ is a parametric representation of the continuous function to be estimated. Again, since $\underline{p}$ is effectively infinite dimensional while the number of observations is limited, the system is underdetermined, i.e., $n \gg m$. The system is generally ill-conditioned as well, so that there are infinite number of unstable solutions (i.e. solutions with large error variance) that satisfy equation (4.14) identically with $\underline{v}$ set to zero. One thus

faces the problem of nonuniqueness compounded with instability.

Before going any further, let us first transform equation (4.14) to

$$\underline{y}' = \underline{F}'\underline{p}' + \underline{v}', \tag{4.15}$$

where $\underline{y}'=\underline{W}^{1/2}\underline{y}^*$, $\underline{p}'=\underline{S}^{1/2}\underline{p}$, $\underline{v}'=W^{1/2}\underline{v}$, and $\underline{F}'=\underline{W}^{1/2}\underline{F}\ \underline{S}^{-1/2}$. The scaling by $\underline{W}^{1/2}$ is necessary because some obsevations may be less reliable. The scaling by $\underline{S}^{1/2}$ is also necessary, because without this scaling, the large weighting coefficients in $\underline{F}$ would tend to put large amplitudes to the associated parameters in an underdetermined system. Both $\underline{S}$ and $\underline{W}$ are symmetric positive definite matrices. Formally, the solution $\underline{p}'$ of (4.15) can be expressed as a weighted sum of normalized orthogonal vectors $\underline{\nu}_j$ belonging to a complete set such that

$$\underline{p}' = \sum_{j=1}^{n} a_j\ \underline{\nu}_j. \tag{4.16}$$

The a_j's are the unknown coefficients which we hope to determine from the data and from some criteria in the inverse problem. Choosing the right set of $\underline{\nu}_j$'s is crucial to the success of the inversion.

To deal with nonuniqueness and instability, linear inverse methods proceed with a spectral decomposition of $\underline{F}'$, that is the

singular value decomposition (SVD) of $\underline{F}'$ (Lanczos, 1961). The SVD gives

$$\underline{F}' = \underline{U}\,\underline{A}\,\underline{V}^T , \tag{4.17}$$

where $\underline{A}$ is an mxm diagonal matrix with nonnegative elements, and $\underline{U}$ and $\underline{V}$ are mxm and nxm matrices, respectively. The jth diagonal element (at the jth row and jth column) of $\underline{A}$ is the jth singular value λ_j or the square root of the jth eigenvalue λ_j^2 of either one the following eigenvalue-eigenvector problems:

$$\underline{F}'\underline{F}'^T\underline{u}_j = \lambda_j^2\underline{u}_j;\ j=1,2,\ldots,m, \tag{4.18a}$$

or

$$\underline{F}'^T\underline{F}'\underline{v}_j = \lambda_j^2\underline{v}_j;\ j=1,2,\ldots,n, \tag{4.18b}$$

with $\lambda_j>\lambda_{j+1}$ by convention. The solution for the eigenvectors $\underline{v}_j$ of equation (4.18b) is the choice of the set of basis vectors for $\underline{p}'$ in equation (4.16). The jth columns of $\underline{U}$ and $\underline{V}$ are the eigenvectors $\underline{u}_j$ and v_j, respectively. Notice that $\lambda_j=0$ for $j>m$, and the corresponding null-space eigenvectors $\underline{v}_j$'s with $j>m$, even though are constructible, they are not included in $\underline{V}$ in the decomposition of $\underline{F}'$. It is because they are not resolvable or constrained by the data: Any combination of the null-space eigenvectors is a solution to the homogeneous equation $\underline{F}'\underline{p}'=0$, and they are the reason for nonuniqueness. At this stage, a good

strategy is to ignore the null-space completely and accept the unique particular solution $\underline{p}_p'$ as the approximate solution for $\underline{p}'$. By substituting (4.16) and (4.17) into (4.15) with $\underline{v}'$ set to zero, and using the equalities $(\underline{U}\ \underline{A}\ \underline{V}^T)\underline{v}_j=\lambda_j\underline{u}_j$ and $(\underline{U}\ \underline{A}\ \underline{V}^T)^T\underline{u}_j=\lambda_j\underline{v}_j$ and the orthonormality of the eigenvectors, we obtain $a_j=\lambda_j^{-1}(\underline{u}_j^T\underline{y}')$, and hence,

$$\underline{p}'_p = \sum_{j=1}^{m} \lambda_j^{-1}(\underline{u}_j^T\underline{y}')\underline{v}_j. \qquad (4.19)$$

When there are less than m independent equations in the system, the number of nonzero singular values (the rank of the system) is actually less than m, and this corresponds to a larger null-space.

Unfortunately, $\underline{p}_p'$ is not a stable solution. As can be seen in equation (4.19), the effect of the noise in $\underline{y}'$ is magnified by the vanishingly small singular values. These appear because of the ill-conditioning of the system, i.e. noise in the model $\underline{F}'$ and almost redundant information in the data. The usefullness of the SVD is now obvious: it provides a meaningful set of basis vectors for $\underline{p}'$, in which the stable and unstable components (vectors) are well distinguished by the sizes of their singular values. Thus, a stable approximate solution $\underline{p}'^*$ can be obtained by discarding or down-weighting the unstable components. A down-weighting technique is to modify equation (4.19) to

$$\underline{p}'^{*} = \sum_{i=1}^{m} \left(\frac{\lambda_j}{\lambda_j^2 + \alpha^{-1}} \right) (\underline{u}_j^T \underline{y}') \underline{v}_j, \tag{4.20}$$

where α is analogous to the Lagrange multiplier of the variational method, representing the signal to noise ratio, and its value is progressively adjusted until the residuals are acceptable and the solution is stable. This is identical to filtering the particular solution by a low-pass filter because the unstable components are usually more oscillatory; indeed, a smoothed version of the particular solution is obtained. This smoothed solution is stable to noise and it is a good approximation when the true solution is also smooth.

Replacing $\underline{p}'^{*}$ with $S^{1/2}\underline{p}^{*}$ in equation (4.20), where $\underline{p}^{*}$ is the estimate to the original parameters $\underline{p}$, and recasting the equation back into matrix form, one obtains

$$\underline{p}^{*} = (\underline{F}^T \underline{W}\, \underline{F} + \underline{S})^{-1} (\underline{F}^T \underline{W}) \underline{y}^{*} \tag{4.21}$$

with α set to unity. Note that one can always make $\alpha=1$ by rescaling $\underline{W}$ and $\underline{S}$. The stochastic analogue of the inverse methods is disclosed by realizing that the linear inverse solution shown in equation (4.21) is actually identical to the MPDE evaluated by minimizing $s(\underline{p})$ of equation (4.7) with $p_0=0$ and $\underline{f}(\underline{p})=\underline{F}\ \underline{p}$, providing that the inverse covariance matrix of noise and the

inverse a priori covariance matrix of $\underline{p}$ are exactly equal to W and S, respectively. Since the MPDE has the statistical property of efficiency (minimum variance) in the case of linear systems, the linear inverse solution becomes the linear minimum variance estimate when $\underline{W}$, $\underline{S}$ and α are identical to $\underline{C}_{\underline{v}}^{-1}$, $\underline{C}_{\underline{p}}^{-1}$ and unity, respectively (Cornuelle, 1983).

It was mentioned earlier that a priori information in the form of a pdf is required to provide uniqueness and/or stability in the stochastic methods when the system is underdetermined and/or ill-conditioned. There are no exceptions in either the variational or inverse methods except that the prior information comes in an equivalent but non-probabilistic form, which is the statement that the continuous function to be estimated is smooth.

We have shown the equivalence of the MPD, variational and inverse methods. Therefore, someone interested only in the final solution and computional efficiency would no doubt formulate the estimation procedure within the context of optimizing objective functions. However, many geophysicists prefer the less efficient but more powerful spectral decomposition technique. Unlike the objective function approaches, in which the information of smoothness is incorporated right at the beginning of and during the optimization process, the spectral decomposition approach does not use this information until the whole spectrum of solutions is

obtained. From there, the resolution of each parameter and the distribution of independent information are simultaneously provided by the spectral decomposition: the jth column of the solution-resolution matrix $\underline{V}\,\underline{V}^T$ indicates how well a delta function located at the jth column of $\underline{p}$ can be resolved, and the jth column of the data-resolution matrix $\underline{U}\,\underline{U}^T$ describes the distribution of the jth independent piece of information in the data. The drawback with spectral expansion techniques is that they are not applicable to systems that are not linear or cannot be linearized.

4.5 Methods for Minimization

In order to focus our attention on the minimization of the objective function $s(\underline{p})$ of equation (4.7), we revise $\underline{p}$ to include both physical parameters and design parameters, and $\underline{y}^*$ to include both experimental data and erroneous design data, when necessary. We would like to emphasize that the minimization of $s(\underline{p})$ is a generalized estimation procedure of many estimation methods. Some widely used numerical techniques of minimization for getting the optimal estimate are reviewed and discussed in this section.

4.5.1 Linear System

The location of the unique minimum of the objective function $s(\underline{p})$ of (4.7) for the linear system in (4.14) can be evaluated, analytically, as

$$\underline{p}^* = \underline{H}^{-1}(\underline{F}^T \underline{C}_v^{-1}\underline{y}^* + \underline{C}_p^{-1}\underline{p}_0), \tag{4.22a}$$

where

$$\underline{H} = (\underline{F}^T\underline{C}_v^{-1}\underline{F} + \underline{C}_p^{-1}) \tag{4.22b}$$

is the Hessian (the matrix of second derivatives) of $s(\underline{p})$. It can be evaluated prior to the finding of $\underline{p}^*$ in a linear system because it is not a function of $\underline{p}^*$. The solution $\underline{p}^*$ exists providing that

the inverse of $\underline{H}$ exists. The most complicated step in solving for $\underline{p}^*$ is, therefore, to invert $\underline{H}$. Gaussian elimation and some of its variations such as the LU and LL^T decompositions which are more convenient for numerical implementations are generally used to perform the task(Dahlquist and Bjorck, 1969).

On the other hand, the problem can also be solved by using the more powerful although less efficient SVD as discussed earlier, so that resolution and information distribution can also be analysed. In order to use the SVD, the eigenvalue-eigenvector problems of (4.18a) and (4.18b) must be attacked. This causes loss of efficiency because finding eigenvalues is a time consuming task. The numerically stable QR algorithm for finding eigenvalues and the inverse iterative methods for evaluating eigenvectors are recommended (Acton, 1970).

4.5.2 Nonlinear System

Numerous methods for minimization have been developed in recent years, but there is no single scheme that works for all problems. A method may work well for one type of objective function but fail for another type. However, most of the methods are iterative in nature, requiring an external initial guess $\underline{p}_1$ for the minimum-point $\underline{p}^*$, and then generating an internal sequence of points at $\underline{p}=\underline{p}_i$ with $i=2,3,\ldots$, progressively, which hopefully converges to $\underline{p}^*$. An iteration is the process of generating a new point in the sequence.

All iterative methods are based on the fundamental reasoning described below.

At the ith iteration, $s(\underline{p})$ near $\underline{p}_i$ may be evaluated as

$$s(\underline{p}_i+\underline{\delta p})=s_i+\underline{g}_i^T\underline{\delta p}+(1/2)\underline{\delta p}^T\underline{H}_i\underline{\delta p}+O(|\underline{\delta p}|^3), \tag{4.23}$$

where $\underline{\delta p}$ is a small vector displacement from $\underline{p}_i$, $s_i=s(p_i)$, and $\underline{g}_i=\underline{g}(\underline{p}_i)$ and $\underline{H}_i=\underline{H}(\underline{p}_i)$ are the gradient vector and Hessian of $s(\underline{p})$ evaluated at p , respectively. Suppose $\underline{p}_i$ is in the quadratic region surrounding $\underline{p}^*$, so that terms of order $\underline{\delta p}^3$ are negligible and $\underline{g}(\underline{p}_i+\underline{\delta p})$ can be expressed as

$$\underline{g}(\underline{p}_i+\underline{\delta p}) = \underline{g}_i + \underline{H}_i\underline{\delta p}. \tag{4.24}$$

Since $\underline{g}=0$ at $\underline{p}^*$, the step (vector displacement) that reaches $\underline{p}^*$ is

$$\underline{\delta p} = -\underline{H}_i^{-1}\underline{g}_i. \tag{4.25}$$

The Newton-Raphson (N-R) method adopts the above scheme explicitly, by setting the ith step $\underline{\delta p}=\underline{p}_{i+1}-\underline{p}_i$ exactly equal to $-\underline{H}_i^{-1}\underline{g}_i$. The N-R method works well for weakly non-linear systems, and in fact it works perfectly in a linear system by requiring only one iteration. Unfortunately, it also fails to work in many cases due to two major weaknesses. First, the method is

mathematically unstable, that is it does not guarantee convergence to a minimum, because an N-R step may not be an acceptable step. An acceptable step is a "down-hill" step such that $s(\underline{p}_{i+1})<s(\underline{p}_i)$; covergence can only be guaranteed if all the steps are acceptable. A down-hill step is ensured if it is taken along a down-hill direction $\underline{d}_i$ such that

$$\underline{d}_i = -\underline{G}_i \underline{g}_i, \tag{4.26}$$

where $\underline{G}_i$ is an arbitrary but positive definite nxn matrix. Realizing that the ith step direction of the N-R method is the one given in equation (4.25) with $\underline{H}_i^{-1}$ replacing $\underline{G}$, and since $\underline{H}_i$ can be nonpositive definite when $\underline{p}_i$ is not inside the quadratic region, stepping up-hill is highly possible along a N-R direction. The second major weakness is that, at each iteration, the method requires the evaluations of $\underline{H}_i$ besides $\underline{g}_i$, and in addition, it also requires the inversion of $\underline{H}$. The analytical expression of $\underline{H}$ as a matrix function of $\underline{p}$ is quite often very difficult to derive, hence the evaluations of $\underline{H}$ at $\underline{p}_i$'s place a heavy burden on the user. This cannot be too pleasing when approximately $n^2/2$ complicated function evaluations are needed at every iteration, not to mention the heavy computational burden of inverting large $\underline{H}_i$ matrices.

In view of the defects of the N-R method, stable and more efficient methods have been sought by many mathematicians. As a consequence, iterative descent gradient methods (gradient methods for short) were developed. These methods abate the burdens on both the user and computer by requiring only the evaluations of $\underline{g}_i$'s but not $\underline{H}_i$'s and $\underline{H}_i^{-1}$'s. More attractively, gradient methods are stable in general cases. Hence, the two major weaknesses of the N-R method disappear in gradient methods.

At each iteration of all the gradient methods, a down-hill direction is selected at the current point and then an acceptable step is taken. This is the reason for their stability. The ith step direction is evaluated by equation (4.26) and $\underline{d}_i$ is always a down-hill direction because the positive definiteness of $\underline{G}_i$ is ensured by the methods. Different gradient methods choose the step directions (or $\underline{G}_i$'s) differently but a similarity of all is that second-derivative information is estimated and incorporated in $\underline{G}_i$ at each step, which gradually evolves to become the inverse Hessian at $\underline{p}^*$ so that equation (4.26) also evolves to become equation (4.25). As a result of this, $\underline{p}^*$ is located. Gradient methods basically fall into two categories: (1) those that require all the $\underline{p}_i$'s to be the minimum-points along $\underline{d}_i$'s, for example, the method of Fletcher and Powell (1963), and (2) those that take acceptable steps but not necessarily reaching the minimum-points along d_i's in all the steps, for example, the Marquardt's method (1963). The trade-off is that the former method requires more

function evaluations per iteration but less iterations while the latter methods require less function evaluations per iteration but more iterations. However, the total number of function evaluations, which determines the efficiency of a method, is usually the same order of magnitude for the two different approaches.

Fletcher and Powell's method (1963) is used in our study. Its use is solely a matter of preference, and we do not claim that it is the best method for the investigation since we have not tested other methods. However, we have found its performance to be more than satisfactory. Although the method requires, at each iteration, to step to the minimum-point along the selected direction, it is still quite efficient because few iterations are needed. It can be shown that if $\underline{p}_i$ is within the quadratic region of a minimum, the method then only requires at most n more iterations to converge to the minimum, where n is the number of unknown parameters of $s(\underline{p})$. The method takes very small steps at the beginning of the minimization process but follows with rapid descent after $\underline{H}(\underline{p}^*)^{-1}$ has been closely approximated by $\underline{G}_i$.

4.6 Error Of The Estimate

An estimate has no meaning by itself since it should not be trusted for the interpretation of the true physical situation without the knowledge of its error. To investigate whether an estimate is well or ill determined, one can either employ nonstatistical response surface techniques (Bard, 1974) or, similarly, the statistical analyses of variance (Jenkins and Watts, 1969, Bard, 1974).

In the response surface technique, we say that there is no reason to prefer the minimum at $\underline{p}^*$ as the solution over any other value of $\underline{p}$ for which

$$s(\underline{p})-s(\underline{p}^*) \sim 1/2\ (\underline{p}-\underline{p}^*)^T \underline{H}(\underline{p}^*)(\underline{p}-\underline{p}^*) < \varepsilon, \tag{4.27}$$

where ε is an arbitrary small constant and $\sim$ is replaced by $=$ when the system is linear, so that the larger (smaller) the diagonal elements of $\underline{H}(\underline{p}^*)$ ($\underline{H}(\underline{p}^*)^{-1}$) are, the better the corresponding parameters are estimated.

On the other hand, statistically, an approximation of the logarithm of the posterior distribution can be expressed as

$$\log[P(\underline{p}\,|\underline{y}^*)] \propto -1/2\ (\underline{p}-\underline{p}^*)^T \underline{H}(\underline{p}^*)(\underline{p}-\underline{p}^*). \tag{4.28}$$

This corresponds to approximating the posterior distribution with a

normal distribution, and the approximation is good near $\underline{p}^*$ if the objective function is symmetric at the minimum. The tails of the distribution are of no concern in the error analyses. Assuming that $\underline{p}^*$ is quite close to the expectation, it then follows that an approximation of the covariance matrix of the error of the estimate is

$$C_{\underline{\Delta p}^*} = \underline{H}(\underline{p}^*)^{-1} \tag{4.29}$$

Thus, the diagonal elements of $\underline{H}(\underline{p}^*)^{-1}$ are approximately the variances of the estimates of the parameters.

It was shown that whether considering statistics or not, $\underline{H}(\underline{p}^*)^{-1}$ is the important measure of error of $\underline{p}^*$. We would like to mention that a problem in design is to pre-arrange the design parameters so that the diagonal elements of $\underline{H}(\underline{p}^*)^{-1}$ are minimized for an expected range of possible values of $\underline{p}^*$. This design problem is easier to tackle if the system is linear since in this case H does not depend on $\underline{p}$.

Since $\underline{H}(\underline{p}^*)$ is not a diagonal matrix in general, the errors of different parameters can be correlated. However, it is of interest in many aspects of error analyses, for example, statistical inference, to look at uncorrelated errors. As a result, a linear transformation

$$\underline{p}''=\underline{Q}^T\underline{p} \tag{4.30}$$

of the original parameters $\underline{p}$ is often exercised so as to bring about uncorrelated errors of the transformed parameters; $\underline{Q}$ is an nxn matrix. In other words, uncorrelated errors of linear combinations of the original parameters are analysed. These orthogonal combinations can be found by a SVD of $\underline{H}(\underline{p}^*)$ or $\underline{H}(\underline{p}^*)^{-1}$. Let us consider the decomposition of $\underline{H}(\underline{p}^*)$ such that

$$\underline{H}(\underline{p}^*) = \underline{Q}\ \underline{D}\ \underline{Q}^T, \tag{4.31}$$

where $\underline{D}$ is a nxn nonnegative definite diagonal matrix consisting of the nonnegative singular values. The substitution of equation (4.26) in equation (4.24) with $\underline{p}''=\underline{Q}^T\underline{p}$ and $\underline{p}''^*=\underline{Q}^T\underline{p}^*$ gives

$$\log[P(\underline{p}''|\underline{y}^*)] \propto -1/2\ (\underline{p}''-\underline{p}''^*)^T\underline{D}(\underline{p}''-\underline{p}''^*), \tag{4.32}$$

where $\underline{p}''^*$ is the estimate of $\underline{p}''$. It is seen that $\underline{D}$ is the covariance matrix of the error of $\underline{p}''^*$, and the errors of these transformed parameters, which are linear combinations of the original parameters, are not correlated because $\underline{D}$ is diagonal.

In a nonlinear system, since the posterior distribution may not be unimodal, many initial guesses are required in the minimization procedure inorder to expose the least minimum or to see if all of them converge to the same $\underline{p}^*$. If more than one global minimum is found, the estimation problem is nonunique.

4.7 Goodness Of A Model

Even more important than judging the reliabilty of $\underline{p}^*$ is judging the reliability of a model. The goodness of a model can be assessed by analysising its final residuals. A model can never be proved correct in principle, but it can be proved incorrect or inconsistent or inferior to other models. Some uniform measures of goodness based on the final residuals can be evaluated for different but plausible models, which can then be compared to discriminate between models and various hypotheses.

If a model is accurate and parameters are well determined, the residuals will reflect the experimental random noise. In fact, residuals are biased estimates of noise: they should be smaller than the actual random error on the average (Bard, 1974). Some of the most common tests on residuals in time series are Chi-square goodness-of-fit, runs and correlation tests (Bendat and Piersol, 1971), which are used to confirm a model by verifing noise statistics such as normality, stationarity and lack of correlation. However, these tests are not applicable when only a few realizations of the same random variable are made, as is usually the case in expensive oceanographic experiments, for example CTD surveys. Fortunately, in model discrimination, there is less interest in knowing how well the residuals of the best model resemble the noise properties than in knowing how well the data are resolved by the best model as compared to the other models; keeping in mind that

some of the noise properties are assumptions anyway.

In what follows, we present some unsophisticated, yet very useful, measures of goodness of model, which are often sufficient to serve the purpose of model discrimination.

The simplest of all measures is the weighted sum of products of the final residuals, that is

$$R = c\ \underline{e}^T \underline{C}_v^{-1} \underline{e}, \tag{4.33}$$

where R is a Chi-square distributed random variable with m degrees of freedom, $\underline{e}$ is the final residual vector, and the adjusting factor c is the total number of experimental, a priori, and erroneous design data divided by the same number less the total number of unknown parameters. If a priori data are available and used in the objective function of equation (4.7), then $c=m+n+r/(m+n+r)-(n+r)$ $=m+n+r/m$. If a priori data are not used, then $c=m+r/(m+r)-(n+r)$ $=m+r/m-n$, where r is the number of erroneous design data or parameters. The factor c is needed to adjust the inverse covariance matrix of noise to equal that of the final residuals due to the bias (Bard, 1974). A significance level can be selected for rejecting models on the two edges of the distribution.

In general, the smaller the misfit between the data and the model, the better the model and the resolution in the solution (or parameter) space are. However, care must be taken when the misfit is extremely small, because we may have an unstable system instead

of a perfect model. Backus and Gilbert (1970) have shown that trade-off between resolution and the statistical reliability of the estimate exists when noise is present. Moreover, for an ill-conditioned system, the variance of the estimate increases without bound as resolution is pushed beyond the limit imposed by the data. Thus, a model is acceptable if and only if both the variance and R, which is a measure of misfit and hence of resolution also, are acceptable.

To illustrate the trade-off between resolution and reliability, consider the linear system (4.15). An estimate may be constructed using (4.19), where the basis vectors $\underline{v}_j$ are weighted and then summed to give the estimate. Since the weighting on $\underline{v}_j$ is the product of the inverse singular value λ_j^{-1} and the projection of the observations $\underline{y}'$ onto the corresponding eigenvector $\underline{u}_j$ in the data space, the small λ_j's can translate the experimental noise into large estimation errors. Clearly then, the reliability of the estimate can only be improved by degrading the resolution, that is discarding or down-weighting the $\underline{v}_j$'s that have small λ_j.

There are two other measures which are often used to judge the success of a model in predicting (interpolating) data. They are the correlation coefficient between observed and predicted signal

$$C = \frac{\underline{y}^{*T}\underline{C}_{\underline{v}}^{-1}\underline{f}(\underline{x}^*,\underline{p}^*)}{(\underline{y}^{*T}\underline{C}_{\underline{v}}^{-1}\underline{y}^*)^{1/2}(\underline{f}^T(\underline{x}^*,\underline{p}^*)\underline{C}_{\underline{v}}^{-1}\underline{f}(\underline{x}^*,\underline{p}^*)^{-1/2}} \qquad (4.34)$$

and the amount of signal energy resolved by the prediction

$$E = \frac{\underline{f}^T(\underline{x}^*,\underline{p}^*)\underline{C}_{\underline{y}}^{-1}\underline{f}(\underline{x}^*,\underline{p}^*)}{\underline{y}^{*T}\underline{C}_{\underline{y}}^{-1}\underline{y}^*} \times 100 \text{ percent.} \qquad (4.35)$$

The larger C and E are, the better the model fits the data, but again, these two measurements can be misleading in the case of instability.

The similarities in shape and amplitude between the observed signal and the model prediction are measured by C and E, respectively. In general, C and E are independent, but for a least squares minimization, 100C equals $E^{1/2}$ for the total set of data points. However, for individual subsets of data, C and E remain useful separate pieces of information.

CHAPTER 5

ESTIMATION OF WAVE PARAMETERS AND WAVE DYNAMICS (1): METHOD AND RESULTS

5.1 The Estimator

Our parameter-estimation problem can be phrased as the inversion of the sound-speed perturbations δc based on the data, and contrained by the dynamics of narrow-band planetary waves. A consequence of the addition of the dynamical constraint on δc is the modification of the system to be inverted from highly underdetermined to highly overdetermined. For a small number of waves, the system can be well-conditioned as well. It was the expectation of a small number of waves and of a well-conditioned system that led us to use the MLE estimator instead of the MPD estimator. It was learnt in the estimation process that as the number of waves W increases, the condition of the system deteriorated. However, this has no effect on our investigation, because the optimal wave fit, corresponding to W=3, was unique and well-determined.

With reference to the discussions in Sec. 4.3, the MLE estimator can be formulated as the minimization of an objective function (i.e. likelihood function). Treating the 130 modal-amplitude data $(a_j^o;\ j=1,\ldots,130)$, the 7 time series of sound-speed perturbations $(\delta c_{jk}^o=\delta c_j^o(t=3k \text{ days});\ k=0,\ldots,26$ and

$j=1,\ldots,7$) and the 58 time records of travel-time perturbations ($\delta t^o_{jk}=\delta t^o_j(t=9k \text{ days})$; $k=0,\ldots,8$ and $j=1,\ldots,58$) as 3 independent data subsets (refer to Table 3.2 for the sources of data), and further treating the uncertainty in the nominal horizontal positions, $\underline{\Delta x}$, and the unknown horizontal displacements ($\underline{\delta x_k}=\underline{\delta x}(t=9k \text{ days})$; $k=0,\ldots,8$) of the acoustic moorings as errors in the design parameters, the objective function can be cast as a sum of 5 constraining functions of similar forms of weighted sum of square of residuals. Because there were 9 acoustic moorings, $\underline{\Delta x}$ and $\underline{\delta x_k}$ are 18-dimensional vectors, and we denote their jth components by Δx_j and δx_{jk}, respectively. The objective function can thus be expressed as

$$s(\underline{p},\underline{\Delta x},\underline{\delta x_0},\ldots,\underline{\delta x_8}) = s_a(\underline{p})+s_{\delta c}(\underline{p})+s_{\delta t}(\underline{p},\underline{\Delta x},\underline{\delta x_0},\ldots,\underline{\delta x_8}) + s_{\Delta x}(\underline{\Delta x})+s_{\delta x}(\underline{\delta x_0},\ldots,\underline{\delta x_8}) \qquad (5.1a)$$

with

$$s_a = 1/2 \sum_{j=1}^{130} \sigma_{a,j}^{-2}[a_j^o-a_j^m(\underline{p})]^2, \qquad (5.1b)$$

$$s_{\delta c} = 1/2 \sum_{j=1}^{7} \sum_{k=0}^{26} \sigma_{\delta c,jk}^{-2}[\delta c_{jk}^o-\delta c_{jk}^m(\underline{p})]^2, \qquad (5.1c)$$

and

$$s_{\delta t} = 1/2 \sum_{j=1}^{58} \sum_{k=0}^{8} \sigma_{\delta t,jk}^{-2} [\delta t_{jk}^{o} - \delta t_{jk}^{m}(\underline{p}, \underline{\Delta x}, \underline{\delta x}_{k})]^2 \quad (5.1d)$$

representing the constraints imposed by each of the data subsets on a wave-propagation model (Model 0, 1 or 2) that is characterized by a corresponding set of unknown parameters $\underline{p}$ as described in Sec. 3.4, where a_j^m, δc_{jk}^m and δt_{jk}^m are defined in (3.19), (3.20) and (3.24), respectively. Furthermore,

$$s_{\Delta x} = 1/2 \sum_{j=1}^{18} \sigma_{\Delta x,j}^{-2} \Delta x_j^2 \quad (5.1e)$$

and

$$s_{\delta x} = 1/2 \sum_{j=1}^{18} \sum_{k=0}^{8} \sigma_{\delta x,jk}^{-2} \delta x_{jk}^2 \quad (5.1f)$$

represent the constraints imposed by the erroneous design data on the incorrect horizontal mooring positions. In writing down (5.1b) to (5.1f), we have assumed uncorrelated experimental noise and design-parameter errors, with the variances of a_j^o, δc_{jk}^o, δt_{jk}^o, Δx_j and δx_{jk} being denoted by $\sigma_{a,j}^2$, $\sigma_{\delta c,jk}^2$, $\sigma_{\delta t,jk}^2$, $\sigma_{\Delta x,j}^2$, and $\sigma_{\delta x,j}^2$, respectively. If a priori information on $\underline{p}$ were incorporated in the estimator, (5.1a) would have an additional constraining function, again of a similar

form. The minimum point would then be the MPD estimate and the variance of the estimate should be reduced.

Although a priori information was not incorporated explicitly in the estimator, it was utilized in many related occasions. An implicit usage was in the filtering and reductionof the data (Sec. 3.3). On the other hand, the optimization orminimization of (5.1) was facilitated by reasonable initial guesses of $\underline{p}$ that are consistent with the prior information, for example, the guessed wavelengths are of order 100 km and the guessed wave amplitudes are of order meters per second.

5.2 Assignment Of Noise Variances

In almost any parameter-estimation problem, the variances of measurement noise are generally fairly accurately known. On the other hand, one usually has less idea or no idea at all of what the variances of the model noise might be, especially when the estimation problem corresponds to model identification. However, this inexact knowledge of noise statistics does not in general introduce any major obstacle in solving estimation problems. There are two reasons for this: first, most estimators are not sensitive to slight variations in noise variances, thus as long as the assigned variances are within reasonable ranges of the true variances, the estimate will not be greatly affected. Second, all estimators also generate an estimate of noise, besides an estimate of the parameters, so that one can rely on the noise estimates themselves, that is the final residuals, for refinement of the assigned variances in an iterative estimation process when necessary. The assignment of the noise variances in (5.1) is described below. The assigned values were later found to be consistent with the final residuals, i.e. the final residuals are not consistently larger or smaller than the assigned standard deviations.

By analysing numerous sound-speed profiles acquired by Piips (1967) between Bermuda and Eleuthera, Mooers (1974) found strong evidence for the existence of a first baroclinic semidiurnal tide

with an amplitude of 0.7 m/s in δc at 550 m depth. Furthermore, nonlinear and higher-mode perturbations are neglected in all 3 wave-propagation models (Table 3.3). These neglected higher-order perturbations combined with the internal tide are probably the major contributors to model error. While the errors in the modal-amplitude data are most sensitive to the internal tide due to the lack of temporal filtering, the errors in the filtered δc time records are most sensitive to higher-mode fluctuations. In addition, the δc time records are also subject to errors caused by vertical mooring motion. We guess that the $\sigma_{\delta c,jk}$'s and $\sigma_{a,j}$'s should be roughly 1 m/s, and thus have set $\sigma_{a,j}=\sigma_{\delta c,jk}=1$ m/s for all j and k.

Considering the measurement noise and internal waves and tides alone, Cornuelle et al. (1985) have estimated the daily mean variance of travel-time noise to be 3.6 ms^2. In order to include the errors introduced by the neglected higher-mode perturbations and current effects, and the assumption of travel-time linearity in our models, we have added 5^2 ms^2 to their estimated variance, that is we have made $\sigma^2_{\delta t,jk}=3.6+25$ ms^2. We note that some of the travel times were missing or not resolvable from the 58 ray paths used on some particular days, (especially, during the later period,) and in these cases we set the corresponding variances to infinity.

The available tracking data indicate that the horizontal mooring displacements were of order 200 m. Therefore, we have set $\sigma_{\delta x,jk}=200$ m and 20 m for the untracked and tracked displacements,

respectively. The 20 m standard deviation represents the measurement error expected from the navigation systems. We have further set $\sigma_{\Delta x,j}$=500 m for all j, which is a reasonable value as indicated by the observed travel-time perturbations.

5.3 Results

The iterative descent gradient method of Fletcher and Powell (1963) was used for the wave fits, that is the optimization of (5.1) with different wave-propagation models and numbers of waves. In each minimization, after accepting an initial guess of the unknown parameters $\underline{u}=(\underline{p},\underline{\Delta x},\underline{\delta x}_0,\ldots,\underline{\delta x}_8)$, the method then proceeds to locate a minimum by estimating, progressively, the inverse Hessian matrix $\underline{H}^{*-1}$ (i.e. the inverse of the matrix containing the second derivatives) of $s(\underline{u})$ at the minimum point $\underline{u}^*$ (Sec. 4.5.2). Thus an estimate of $\underline{u}$, $\underline{u}^*$, and an estimate of the error-covariance matrix of $\underline{u}^*$, $\underline{H}^{*-1}$, are generated, simultaneously.

For each of the three models, one to five waves were fitted to the data. At least four different initial guesses of $\underline{p}$ for a given model (Model 0,1 of 2) and number of waves (W=1,2,3,4 or 5) were used in the optimizations to explore the least minimum (i.e. the solution) and to investigate nonuniqueness. All the initial guesses of $\underline{\Delta x}$ and $\underline{\delta x}_k$ were null vectors. While the wave fits with $W\leq 3$ are unique, those with $W>3$ are not. In each fitting with $W\leq 3$, most of the initial guesses converged to the same stationary point where the least minimum occurs, and although a few initial guesses converged to different stationary points, the corresponding minima are considerably larger. For each of the wave fits with $W>3$, different initial guesses resulted in different minima of approximately the same size, hence a unique least minimum could not be identified.

The change from uniqueness to nonuniqueness as W increases is a demonstration of the trade-off between resolution and stability. As W increases, so do the magnitudes of the wavenumber estimates. Thus, finer-scale structures of the perturbations are intended to be resolved with a larger W, but because of the inadequacy of the data in resolving them, the system for δc is rendered underdetermined.

In Cornuelle's (1983) time-independent inversions, no dynamical constraint is imposed on the solution for δc, and in order to ensure uniqueness, he incorporates an a priori covariance of δc that is assumed to be horizontally Gaussian with a decay scale of 100 km in the estimator. This is the same as requiring the solution to be smooth in space. Cornuelle points out that the solution for δc is not sensitive to small variations in the assumed spatial decay scale. We have encountered a similar situation in our time-dependent inversions. An interesting fact is that although the wave-parameter estimates are nonunique in the cases of W=4 and 5, the solution for the corresponding δc is unique. That is, the estimated fields of δc, and the amounts of resolved data variance associated with the different stationary points are practically the same. Indeed, the constraints imposed by the wave dynamics are analogous to the criterion of smoothness, the different stationary points are analogous to the variations of the decay scales in space and time, and a time-dependent inversion is not sensitive to small variations of both decay scales.

In order to assess and compare the different wave fits so that the optimal model and W may be identified, a simple measure of goodness, that is the weighted sum of squares of final residuals R, defined in (4.27), was computed for each of the wave fits. When the estimate is close to the true value, the probability distribution of the final residuals is approximately normal (because the noise distribution is normal) and the covariances of the residuals and the noise are approximately proportional to each other (Ch. 4, Sec. 4.3). Thus, R is approximately a Chi-square distributed random variable with m=841 degrees of freedom where m is the number of data, and the 0.01 significance level of the random variable is at R~940. In Fig. 5.1, we plotted R versus W for each model. It is seen that the performance of Models 1 and 2 is much better than that of Model 0. While none of the wave fits of Model 0 passes the 0.01 significance test, the fits with W=3,4 and 5 of Model 1 and 2 are at and beyond the 0.01 significance level. Although Model 1 and 2 perform equally well, the estimated growth rates of the wave amplitudes in Model 2 do not differ significantly from zero and, in fact, their signs are ambiguous because their rms errors are larger than the estimated growth rates themselves. The lack of ability to determine the growth rates is not surprising, however, because (1) resonant interactions should be rare occurences since the forced waves can grow if and only if they satisfy the dispersion relationship, and (2) even if resonance actually ocurrs, the time scale of the growth, in weak-interaction theory, is much longer than

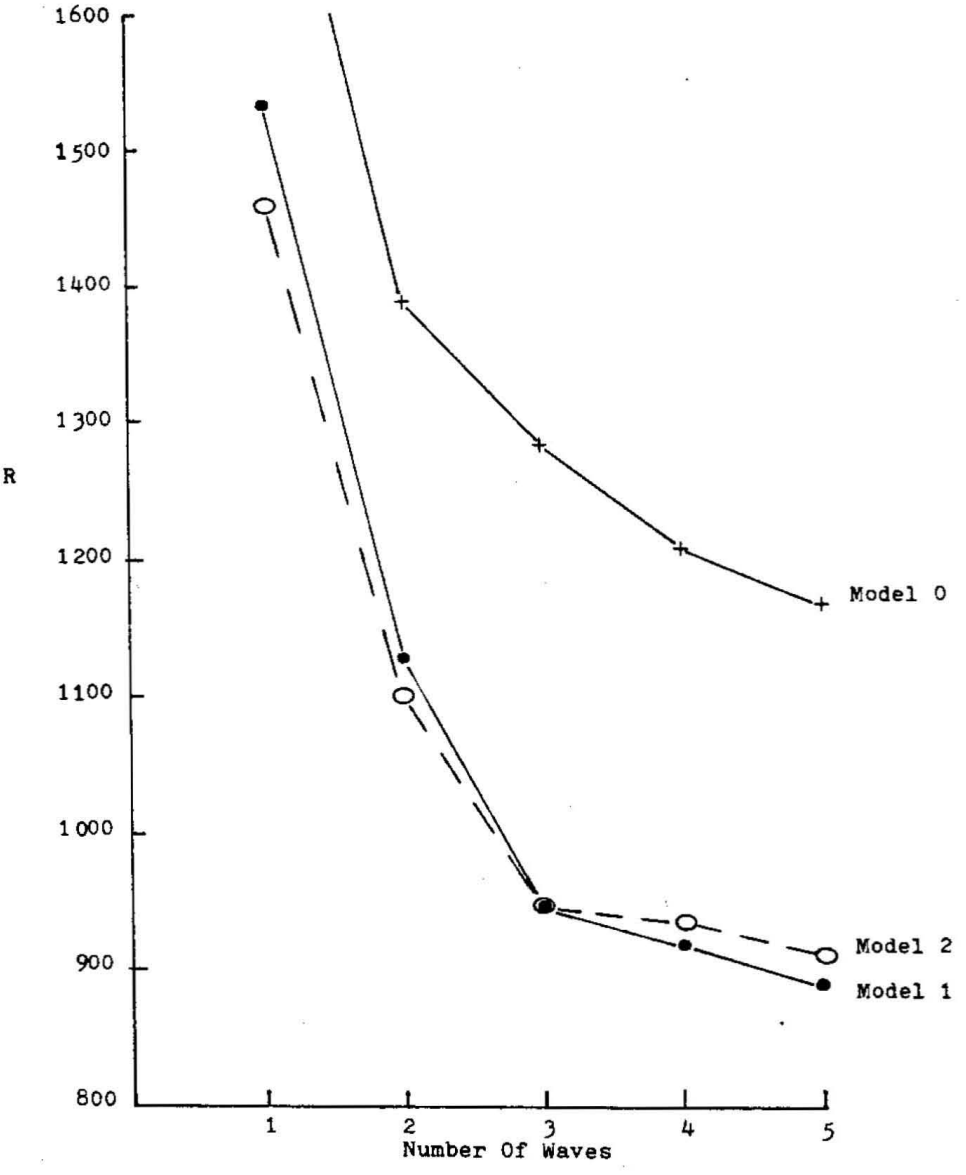

Figure 5.1. Misfits of the 3 wave-propagation models on the data as functions of the number of waves fitted. The measure of misfit is R, which a weighted sum of products of final residuals. The weighting factors are the reciprocals of noise variances.

a wave period and , hence, data measured within a wave period cannot be adequate for observing such phenomena. The reason for fitting Model 2 to the data is to see if there are any surprises that are inconsistent with the theory. For Model 1, the data variance resolved increases by 20 percent as W changes from 2 to 3 and increases by as little as 5 percent as W further changes from 3 to 4 or 5. Also, we must keep in mind that W=4 and 5 correspond to unstable wave fits. Thus, an overall judgement clearly favours Model 1 to be the optimal wave-propagation model in which a mean flow is present and W=3 to be the optimal number of propagating first baroclinic waves.

To make further assessments, we computed for each wave fit the correlation coefficient C_i between the ith independent data subset and the fit, and the amount of variance in the ith data subset resolved by the fit, E_i, using (4.34) and (4.35), where i=1,2 and 3 denote the data subsets of modal amplitudes, δc time records and δt time records, respectively. For Model 1, that is the optimal model, C_i's and E_i's versus W are plotted in Figs. 5.2a and b, respectively. At W=3, i.e., the optimum, we obtain C_i's of 0.8, 0.9 and 0.98 and E_i's of 78, 82 and 96 percent with i=1,2 and 3, respectively. There is no inconsistency although C_3 and E_3 are considerably larger, because a large portion of the variance in the δt time records is resolved by the determination of the mooring-position errors alone. The consistently high correlations and resolutions are a strong evidence of the existence of three

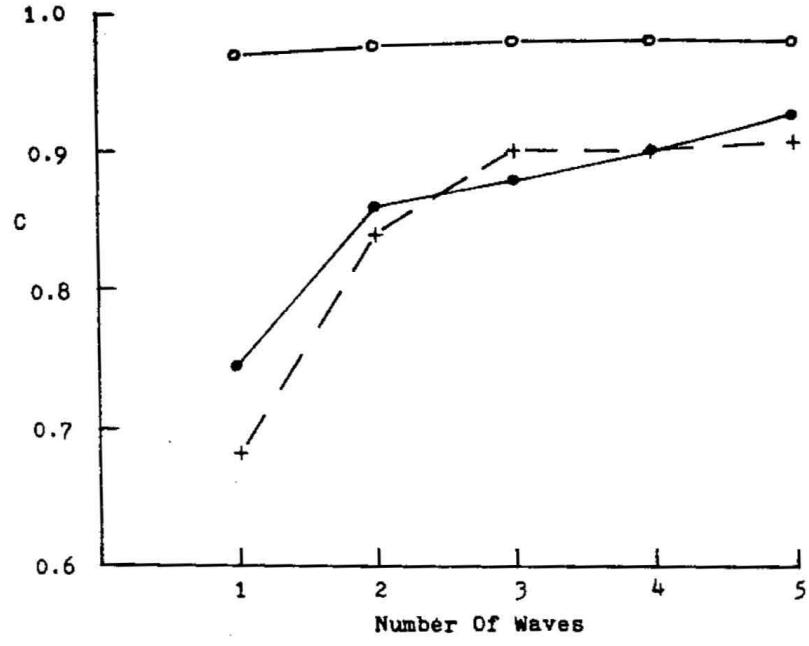

Figure 5.2a. Correlations of the travel-time perturbation records (o), the moored temperature-perturbation records (+) and the modal-amplitude data from CTD casts (.) with the Model-1 fit for a given number of waves.

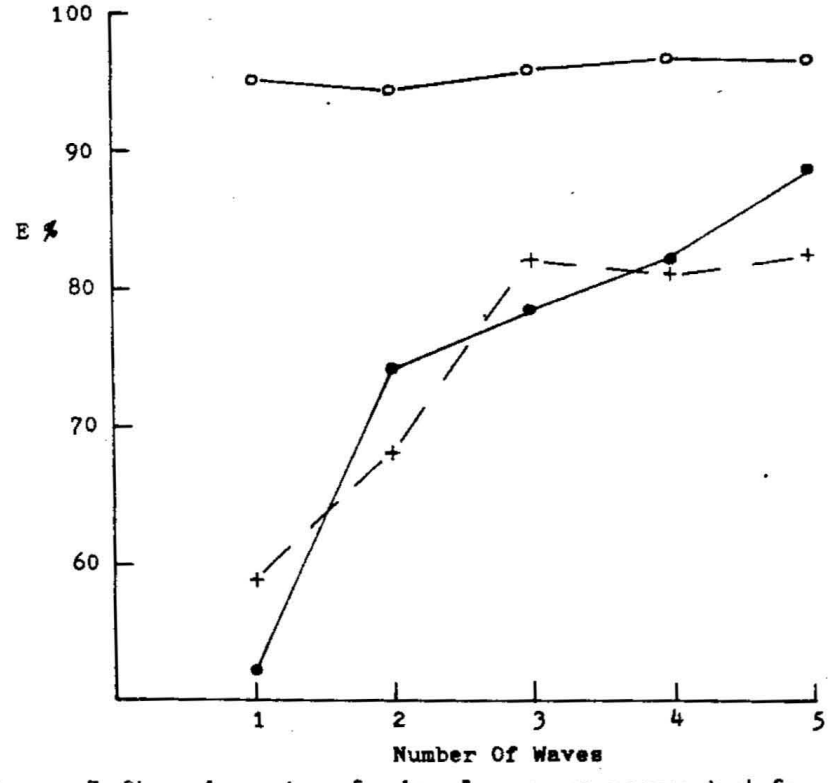

Figure 5.2b. Amounts of signal energy accounted for in the travel-time perturbation records (o), in the moored temperature-perturbation records (+) and in the modal-amplitude data from CTD casts (.) by the Model-1 fit for a given number of waves.

first baroclinic planetary waves in the tomographic region during the experimental period. The optimal values of the parameters for the waves and mean flow, and their standard deviations (square roots of the diagonal elements of $\underline{H}^{*-1}$) are shown in Table 5.1. In the table, the phase and group velocities, the Doppler shifts and the shifted periods themselves, as well as the directions and lengths of the waves are presented. Although the mean flow is very weak, it must be taken into consideration, since it speeds up the phase propagation considerably by generating Doppler effects; it is thus vital to the success of the wave fit.

Table 5.1

The Optimal Estimate Of The Wave Parameters

(a) Independent Wave Parameters; the numbers behind the ± signs are the standard deviations

wave i.d. i.d. no. i	δc amplitude A_i (m/s)	wavenumber vector k_i (1/km)	l_i (1/km)	phase constant γ_i (rad.)
1	1.10±0.13	-.0118±.0011	.0203±.0010	2.13±0.19
2	2.28±0.12	-.0066±.0005	-.0198±.0007	1.51±0.12
3	1.73±0.09	-.0119±.0005	-.0034±.0008	-0.06±0.11

mode no. m	mean-current modal-amplitude vector u_0 (cm/s)	v_0 (cm/s)	zero-reference δc b_0 (m/s)
0	-1.70±0.24	0.11±0.08	
1	-0.76±0.13	0.39±0.09	-1.46±0.21

(b) Dependent Wave Parameters

wave i.d. no. i	wave length (km)	direction of phase (degree)	wave period (days)	Doppler shift period (days)
1	268	120	117	-202
2	300	-108	344	-164
3	509	-164	121	-77

wave i.d. no. i	phase velocities eastward (cm/s)	northward (cm/s)	group velocities eastward (cm/s)	northward (cm/s)
1	-5.25	3.06	-4.24	0.23
2	-3.19	-1.06	-4.30	1.16
3	-5.04	-17.69	-4.23	0.53

The seven observed time records of δc are plotted in Fig. 5.3to 5.9 together with the optimal fits. It is seen that the observations and the optimal interpolations compare favorably. Furthermore, some secondary perturbation with a period of about 20 days superimposed on the primary perturbations created by the 3 linear dispersive waves are found consistently in all the time records. The secondary oscillations were most profound at the mooring site E2, i.e., at (x,y)=(150.7,13.6) km. Because the frequency is below the inertial frequency, this oscillation cannot be due to internal waves; we speculate that the secondary perturbations were caused by the forced waves that oscillate at frequencies equal to the sum of the frequencies of the interacting barotropic and/or baroclinic waves.

To demonstrate that the observed pattern of the fairly complicated system can indeed be reconstructed accurately by the gradual evolution of three waves, we show a time sequence of the estimated and surveyed sound-speed maps at a depth of 700 m in Fig. 5.10 to 5.16. The average sound-speed at that depth is 1506 m/s. The estimated perturbed sound speed on yearday 66, 83, 102, 120 and 137 are contoured in Fig. 5.10, 5.12, 5.13, 5.14 and 5.16, and the observed sound speed from the first and second CTD serveys were contoured in Fig. 5.11 and 5.15, respectively. It is seen that the waves generated a trough that was moving slowly to the west and then produced a front that was advancing from the northeast during the later period.

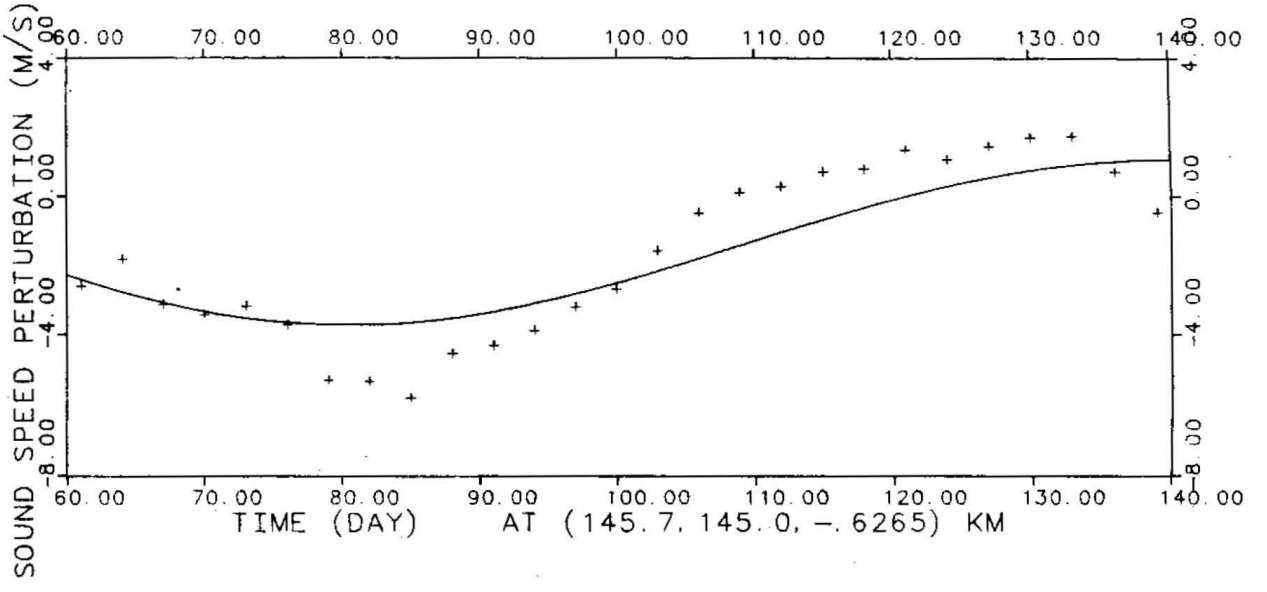

Figure 5.3. Comparison of the optimal wave fit (_____) with the sound-speed perturbation time series (+ + +) observed from the temperature sensor located at $\underline{x}$=(145.7,145.0,-.6265) km.

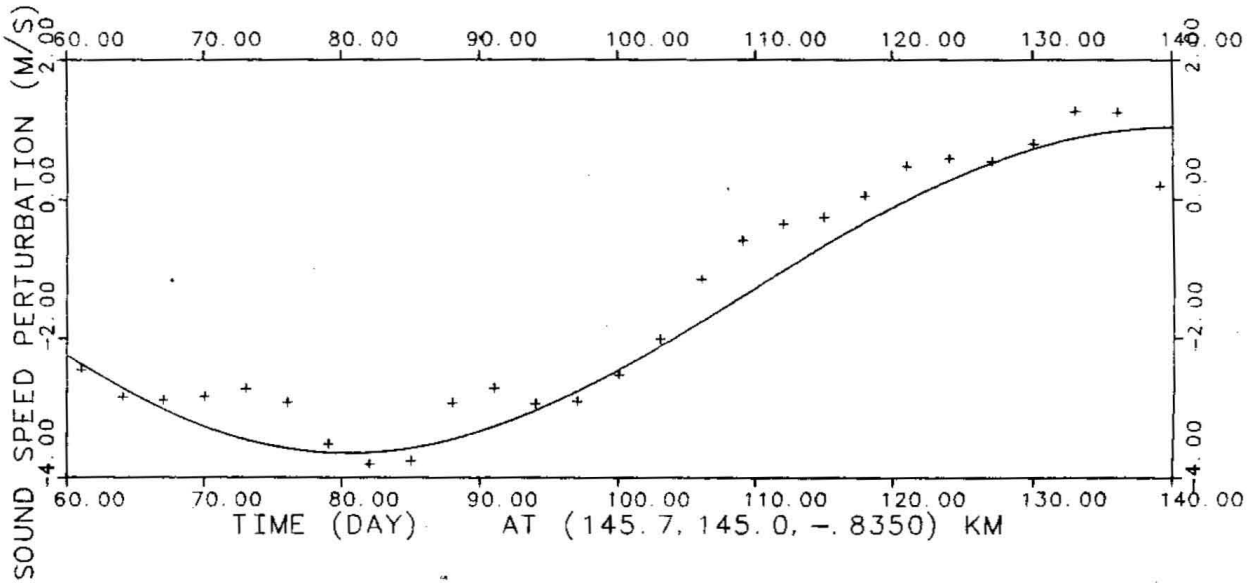

Figure 5.4. Comparison of the optimal wave fit (_____) with the sound-speed perturbation time series (+ + +) observed from the temperature sensor located at $\underline{x}$=(145.7,145.0,-.8350) km.

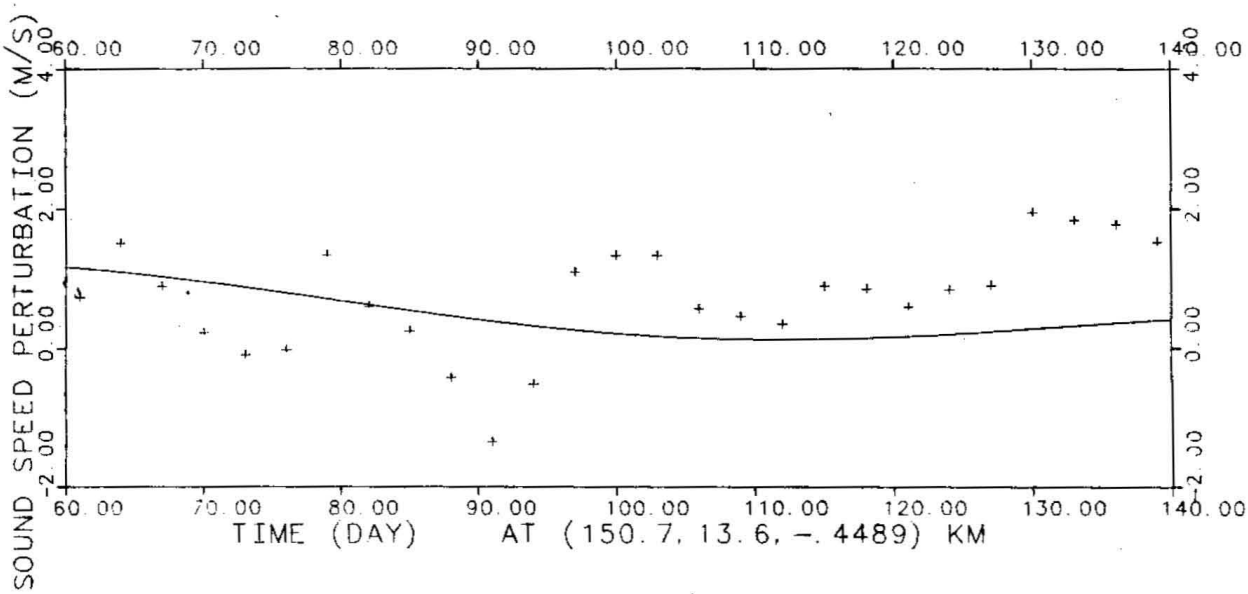

Figure 5.5. Comparison of the optimal wave fit (_____) with the sound-speed perturbation time series (+ + +) observed from the temperature sensor located at $\underline{x}$=(150.7, 13.6,-.4489) km.

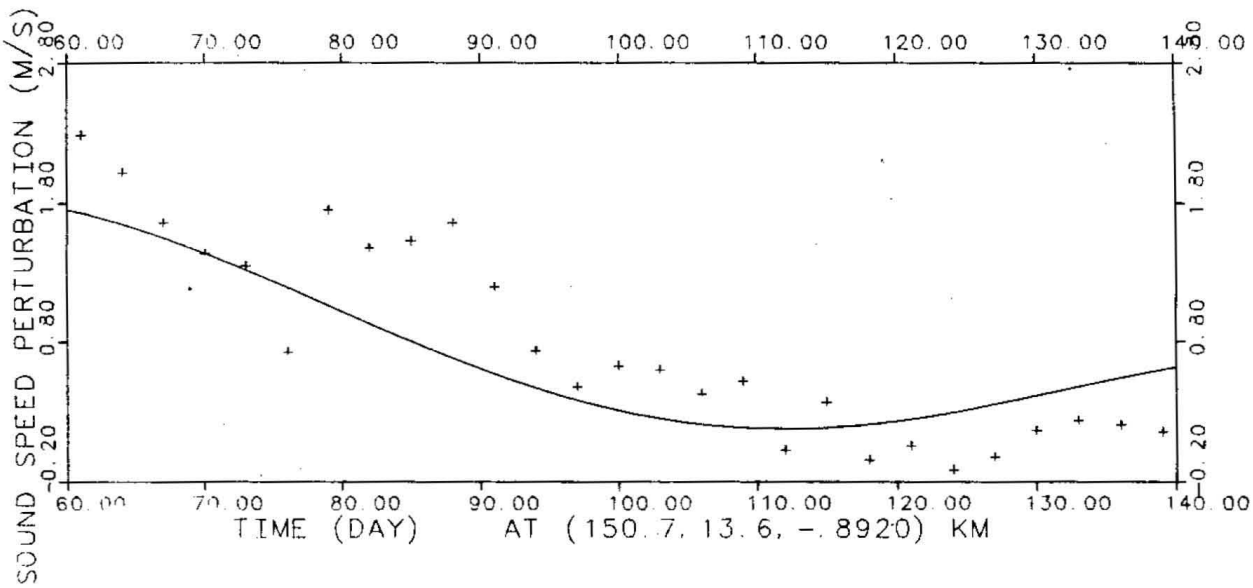

Figure 5.6. Comparison of the optimal wave fit (_____) with the sound-speed perturbation time series (+ + +) observed from the temperature sensor located at $\underline{x}$=(150.7, 13.6,-.8920) km.

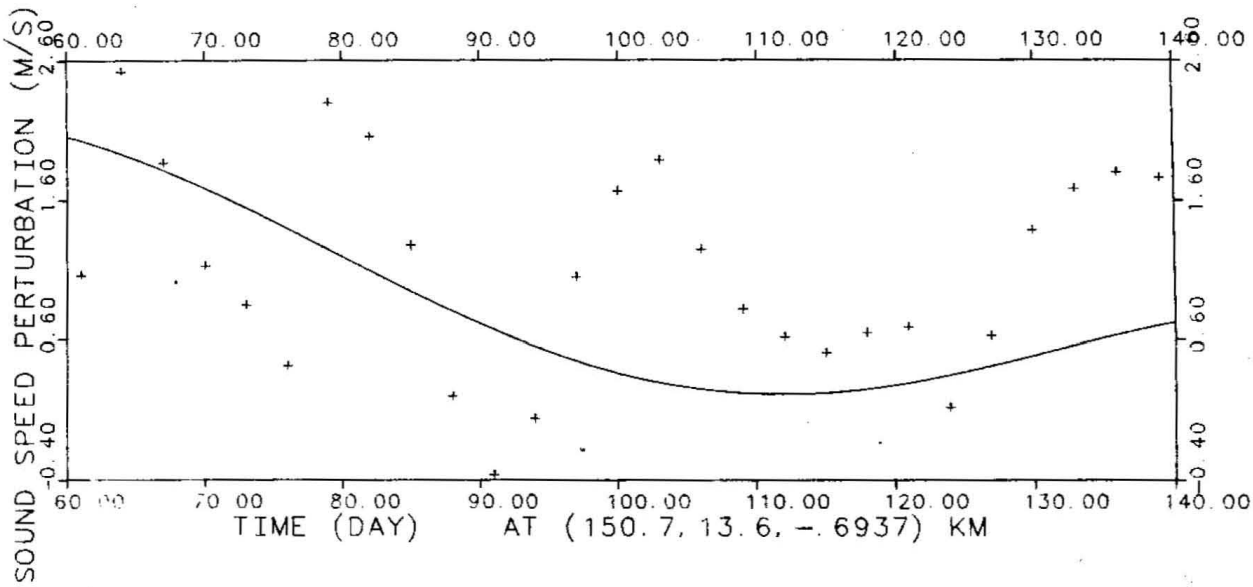

Figure 5.7. Comparison of the optimal wave fit (_____) with the sound-speed perturbation time series (+ + +) observed from the temperature sensor located at $\underline{x}$=(150.7, 13.6,-.6937) km.

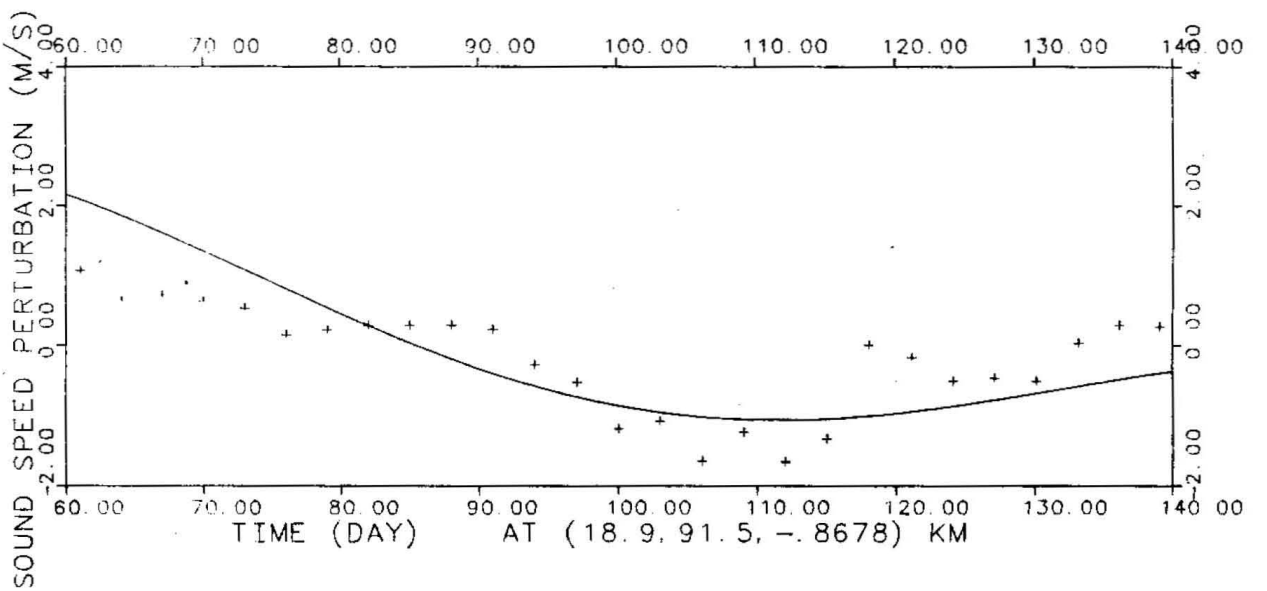

Figure 5.8. Comparison of the optimal wave fit (_____) with the sound-speed perturbation time series (+ + +) observed from the temperature sensor located at $\underline{x}$=(18.9, 91.5,-.8678) km.

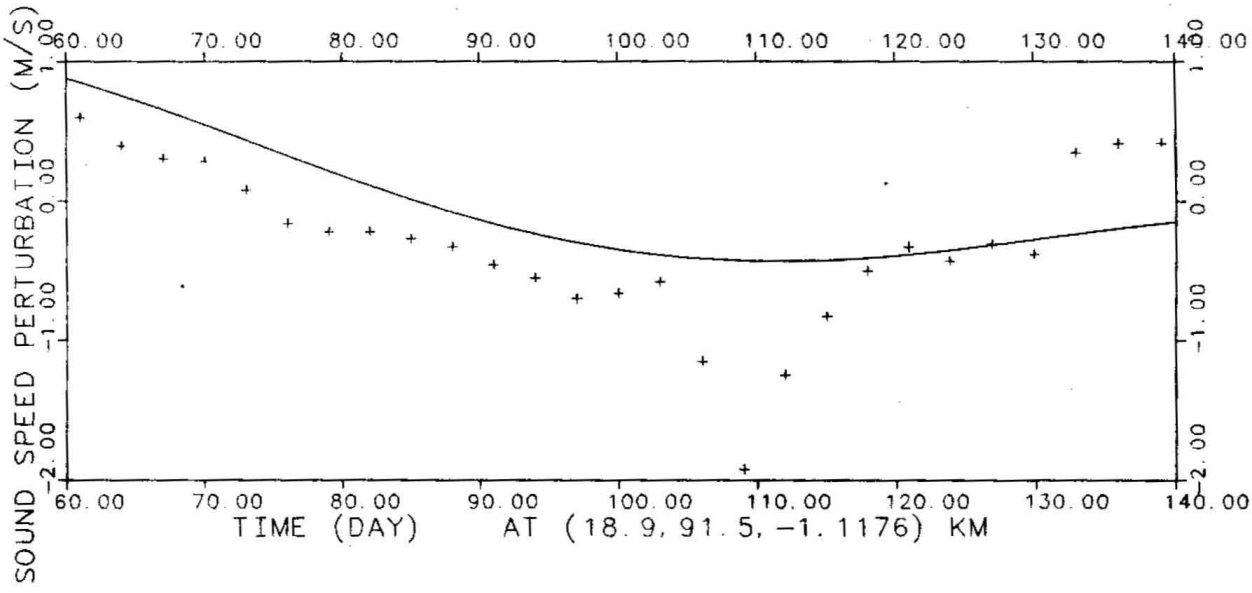

Figure 5.9. Comparison of the optimal wave fit (_____) with the sound-speed perturbation time series (+ + +) observed from the temperature sensor located at $\underline{x}$=(18.9, 91.5,-1.1176) km.

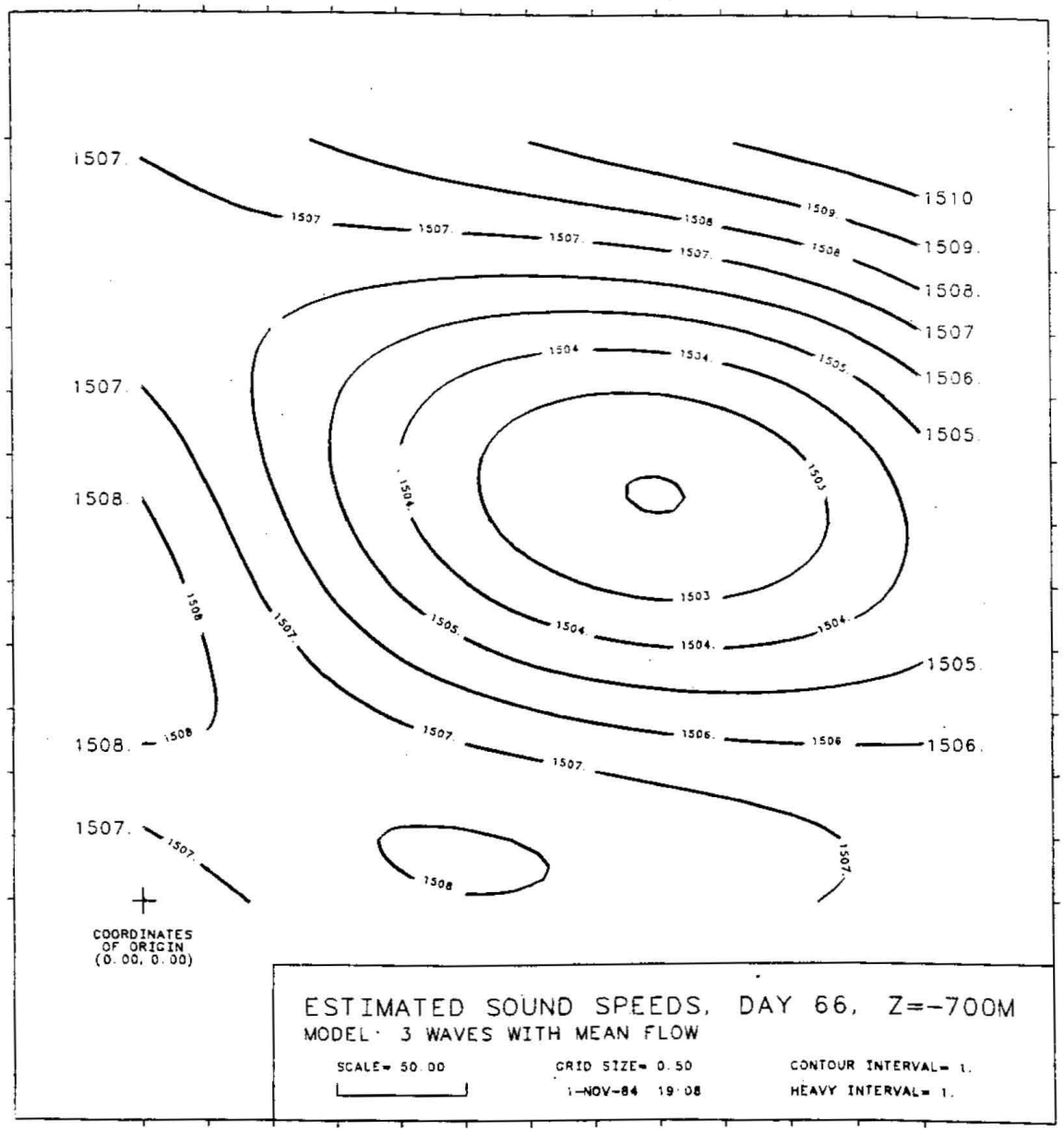

Figure 5.10. Sound-speed map at a depth of 700 m of the optimally estimated wave field in the experimental square on yearday 66. Contour interval is 1 m/s and the reference sound-speed at this depth is 1506 m/s.

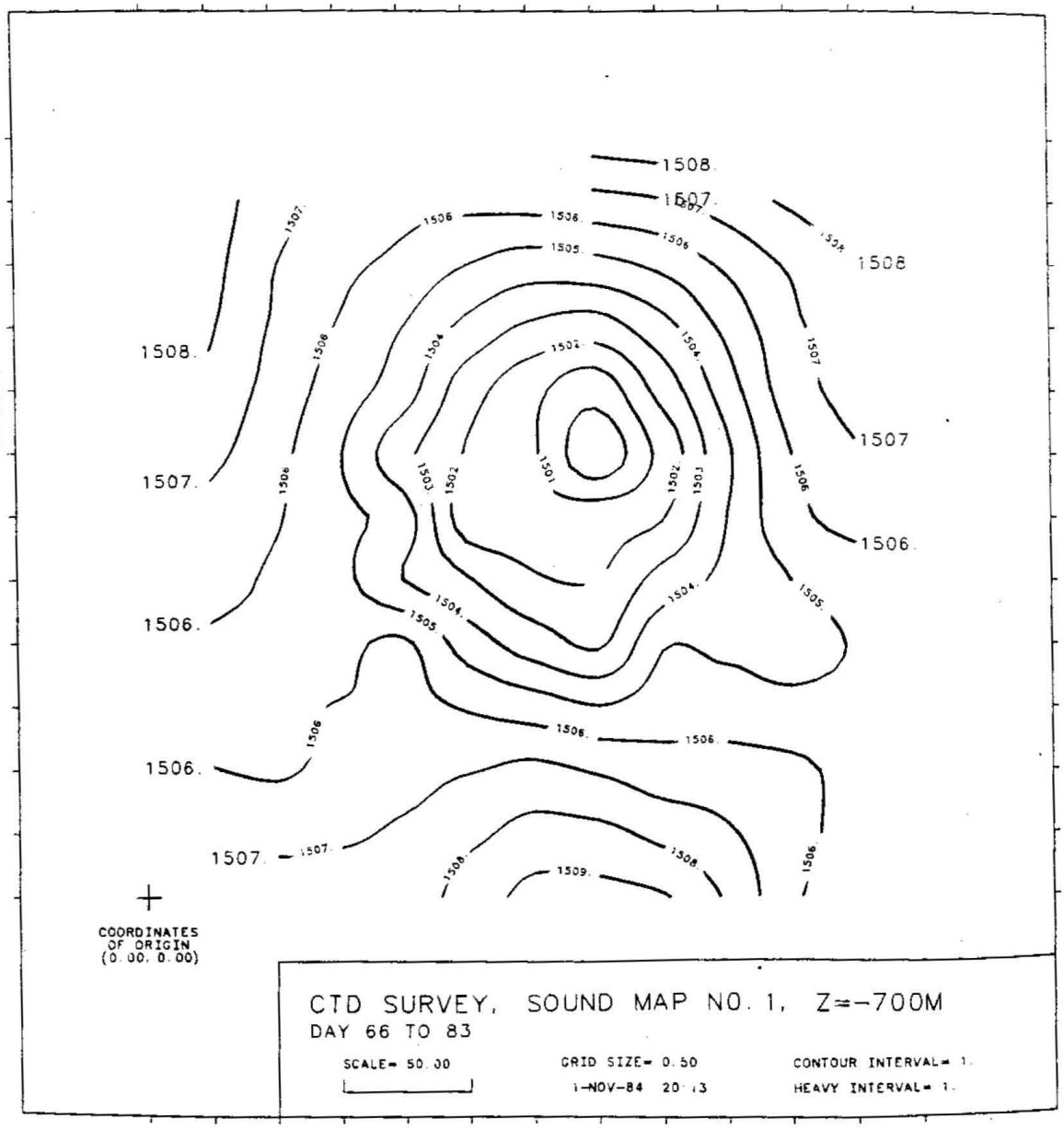

Figure 5.11. Sound-speed at a depth of 700 m in the experimental square, mapped by the 1st CTD survey. The survey began on yearday 66 and ended on yearday 83. Contour interval is 1 m/s and the reference sound-speed at this depth is 1506 m/s.

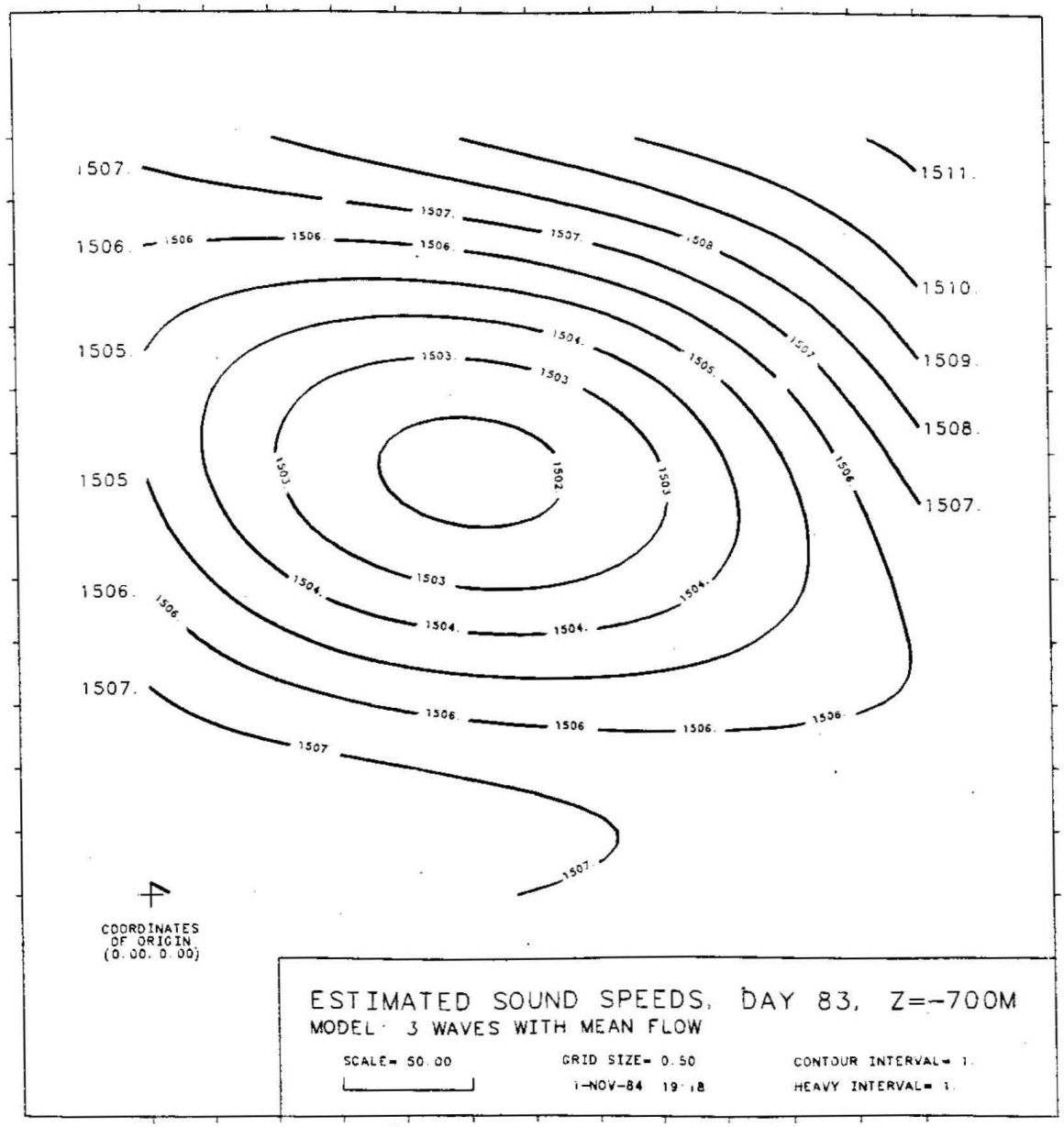

Figure 5.12. Sound-speed map at a depth of 700 m of the optimally estimated wave field in the experimental square on yearday 83. Contour interval is 1 m/s and the reference sound-speed at this depth is 1506 m/s.

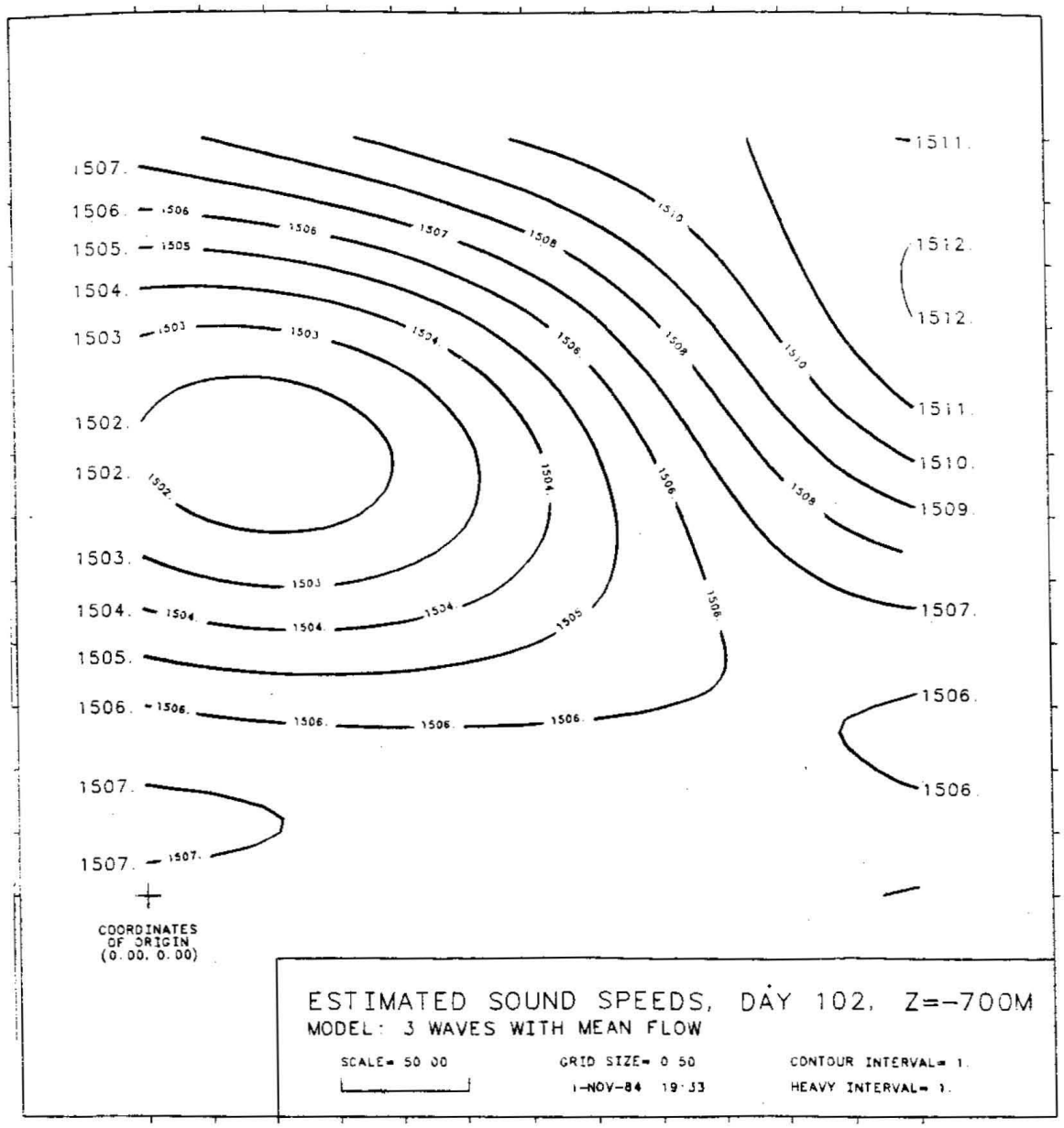

Figure 5.13. Sound-speed map at a depth of 700 m of the optimally estimated wave field in the experimental square on yearday 102. Contour interval is 1 m/s and the reference sound-speed at this depth is 1506 m/s.

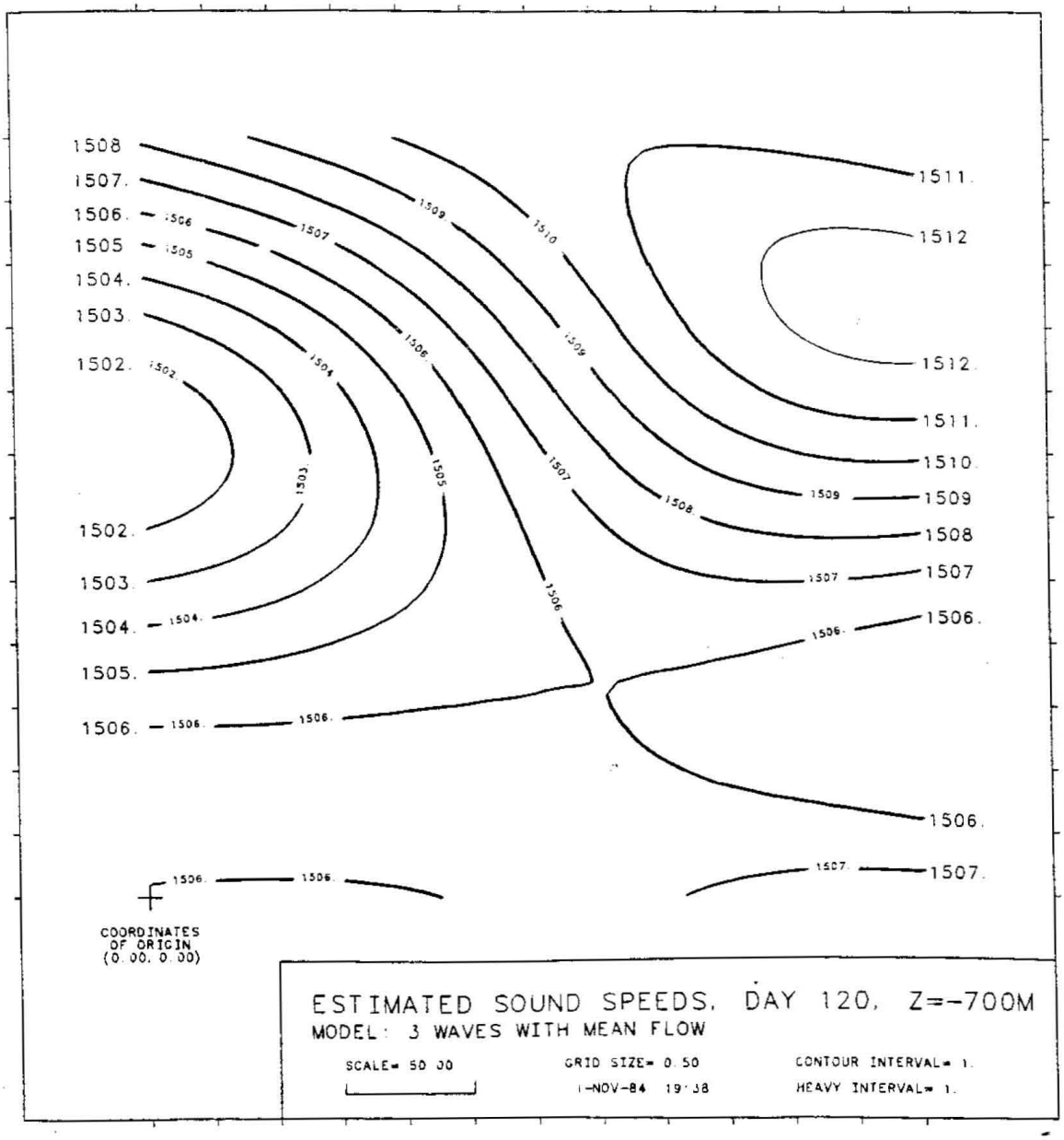

Figure 5.14. Sound-speed map at a depth of 700 m of the optimally estimated wave field in the experimental square on yearday 120. Contour interval is 1 m/s and the reference sound-speed at this depth is 1506 m/s.

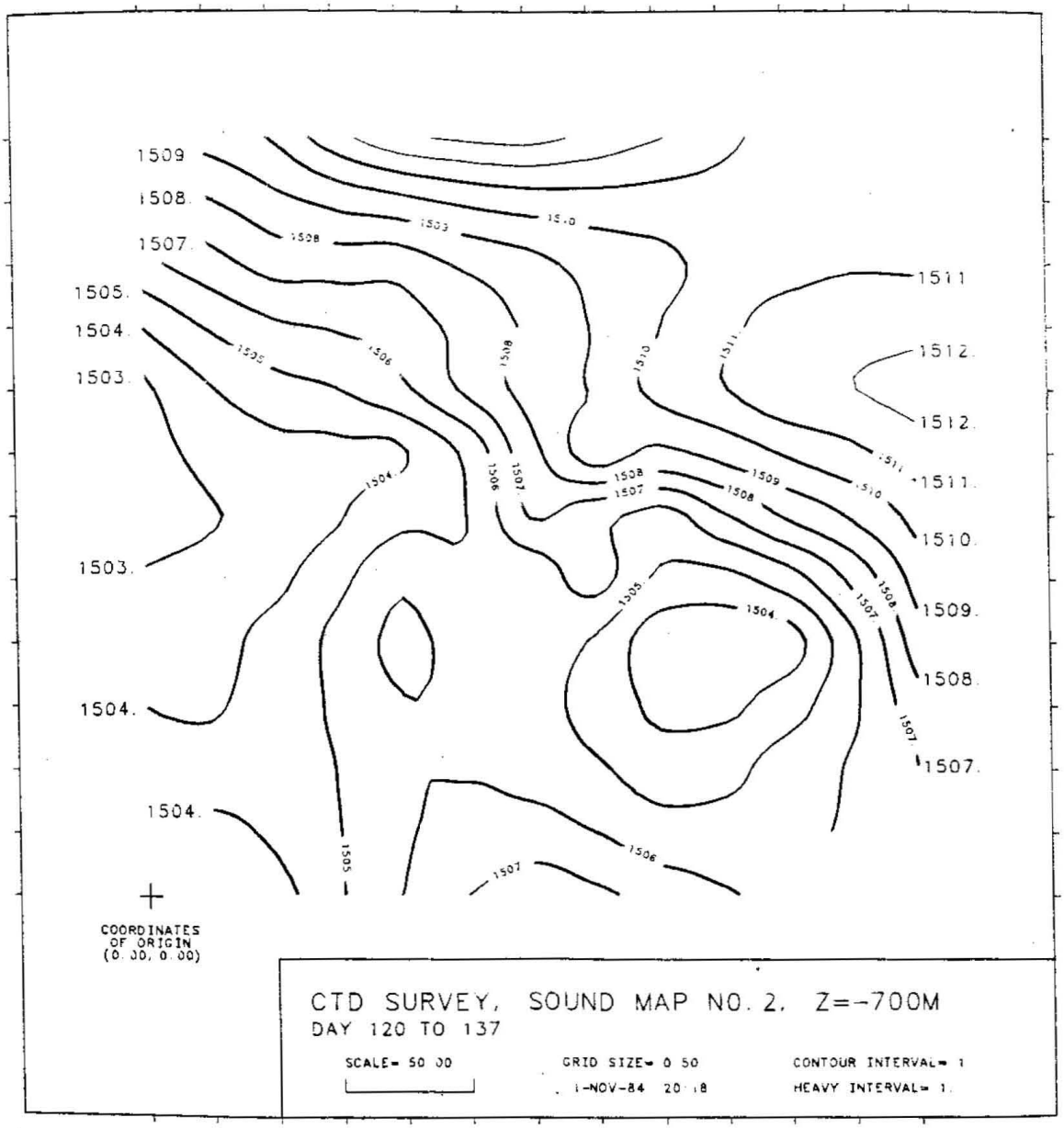

Figure 5.15. Sound-speed at a depth of 700 m in the experimental square, mapped by the 2nd CTD survey. The survey began on yearday 120 and ended on yearday 137. Contour interval is 1 m/s and the reference sound-speed at this depth is 1506 m/s.

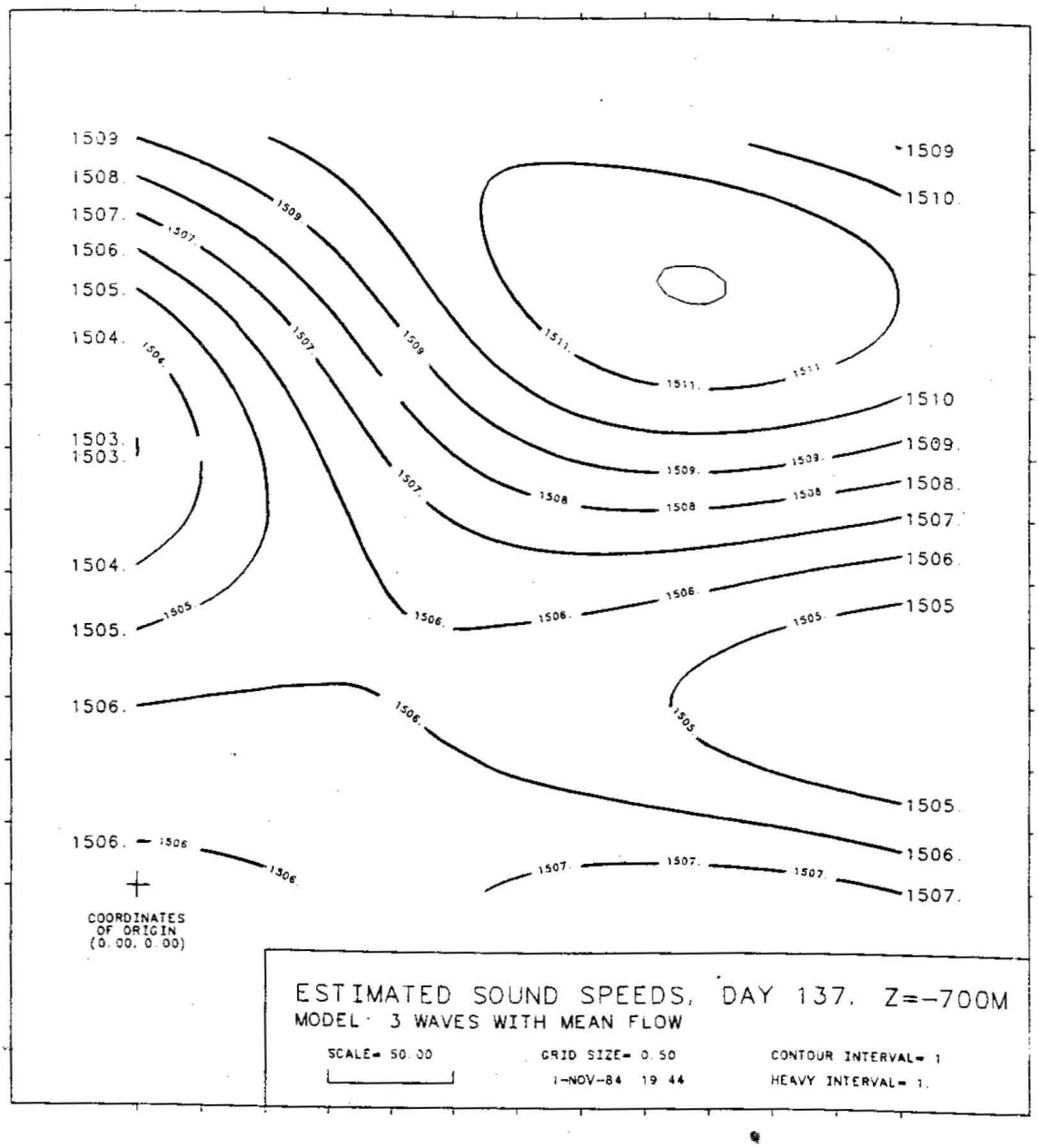

Figure 5.16. Sound-speed map at a depth of 700 m of the optimally estimated wave field in the experimental square on yearday 137. Contour interval is 1 m/s and the reference sound-speed at this depth is 1506 m/s.

The covariance matrix $\underline{H}^{*-1}_{\underline{p}}$ of the wave-parameter estimate $\underline{p}^*$ (i.e. the corresponding block in the inverse Hessian matrix of the objective function evaluated at the minimum point) gives indications on which wave parameters, or linear combinations of wave parameters, are well determined, and which are poorly determined. A simple measure of the quality of the estimate is given by the diagonal elements of $\underline{H}^{*-1}_{\underline{p}}$, which are the variances of the errors in the estimate; the standard deviations are listed in Table 5.1a. However, the presence of nonzero off-diagonal elements implies that the errors are correlated, and a full description of the error structure must take all the elements of the matrix into account. As discussed in Sec. 4.6, a full description may be obtained by finding the eigenvalue decomposition of $\underline{H}^{*-1}_{\underline{p}}$ such that $\underline{H}^{*-1}_{\underline{p}}=\underline{U}\ \underline{D}\ \underline{U}^T$, where $\underline{D}$ is the diagonal matrix containing the eigenvalues and $\underline{U}$ is the matrix containing the eigenvectors in its columns, so that new variables defined by $\underline{p}'=\underline{U}^T\underline{p}$ and representing a set of linear combinations of the wave parameters would have uncorrelated errors in their estimate $\underline{p}'^*=\underline{U}^T\underline{p}^*$. The error variance of $\underline{p}'^*$ is $\underline{D}$. We have performed the decomposition and found the set of linear combinations of wave parameters. We have found that all the 17 linear combinations were well determined. The difference between the variances of the best and the worst determined linear combinations is small. The 17 linear combinations will not be listed since they serve no further purpose in this investigation.

Finally, it is desirable to obtain an error estimate of the estimated sound-speed perturbation $\delta c^* = \delta c^m(\underline{p}^*)$ due to the error $\underline{\Delta p}^*$ of the estimated wave parameters $\underline{p}^*$. Through a linearization of the wave and mean-flow induced sound-speed perturbation $\delta c^m(\underline{p})$ about $\underline{p}^*$, the error of δc^* can be approximated by

$$\Delta\delta c^* \sim \left[\frac{\partial \delta c^m(\underline{p}^*)}{\partial \underline{p}} \right]^T \underline{\Delta p}^*. \tag{5.2}$$

It then follows that the variance of Δc^* can be written as

$$\langle \Delta\delta c^{*2} \rangle \sim \left[\frac{\partial \delta c^m(\underline{p}^*)}{\partial \underline{p}}\right]^T \underline{H}_{\underline{p}}^{*-1} \left[\frac{\partial \delta c^m(\underline{p}^*)}{\partial \underline{p}}\right], \tag{5.3}$$

where $\underline{H}_{\underline{p}}^{*-1}$ is the covariance matrix of $\underline{\Delta p}^*$. In Fig. 5.17, 5.18 and 5.19, we show the contour plots of the standard deviation of δc^* (i.e. the square root of (5.3)) at a depth of 700 m on yeardays 83, 102 and 120, respectively. Because the densities of the ray paths and the CTD stations were much higher in the middle of the area, the errors are smaller there. Furthermore, since there was an enviromental mooring E2 on the southern boundary, the errors near this boundary is smaller than those near the northern boundary where no enviromental mooring was deployed (see Fig. 3.1). The constraint imposed by the wave dynamics had introduced a high correlation

between the sound-speed perturbations at different locations and times; thus the errors in all the maps stay within a pretty narrow range.

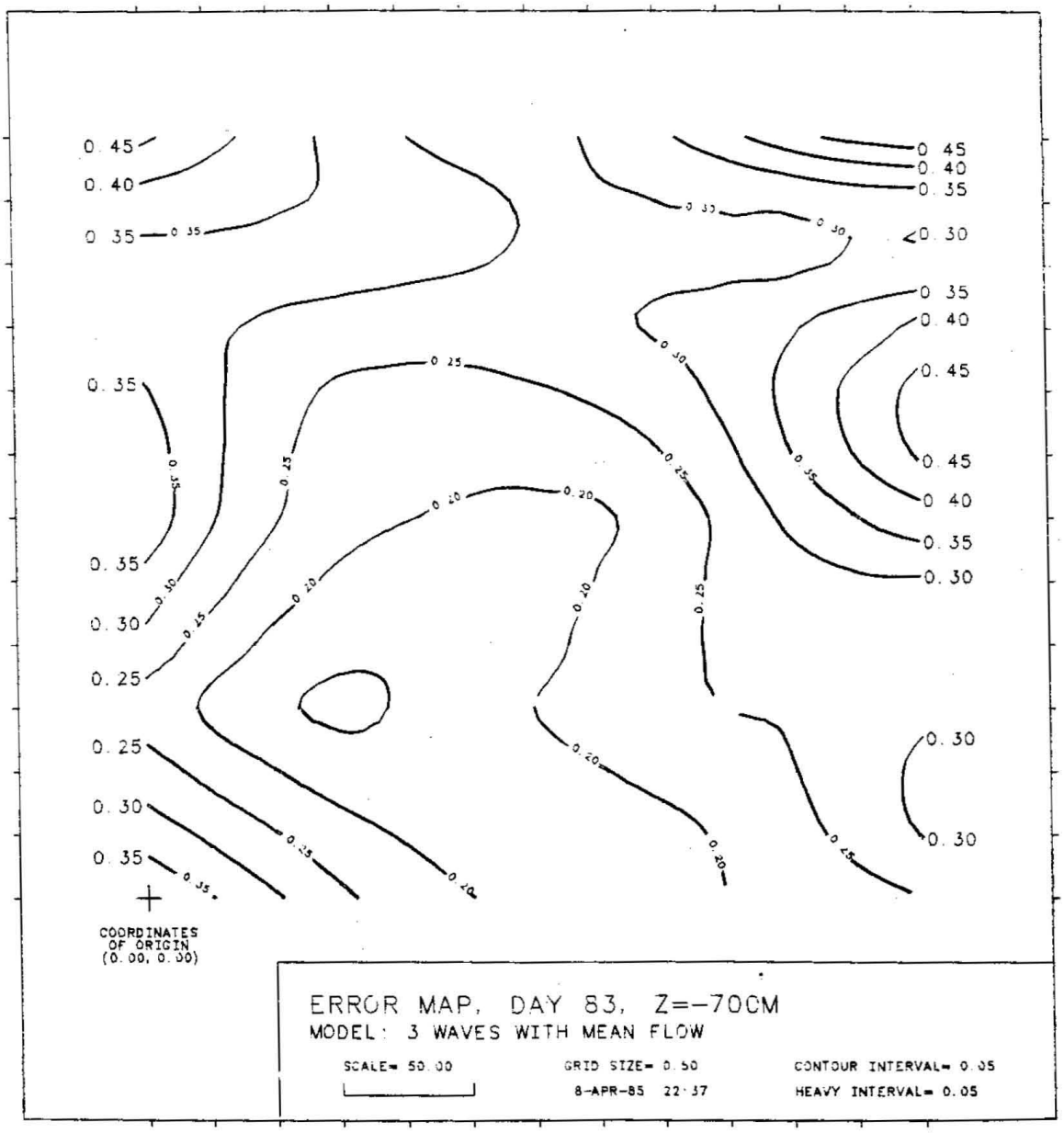

Figure 5.17. Error map, showing contours of the standard deviation at a depth of 700 m of the optimally estimated sound-speed perturbations in the wave field in the experimental square on yearday 83. Contour interval is 0.05 m/s.

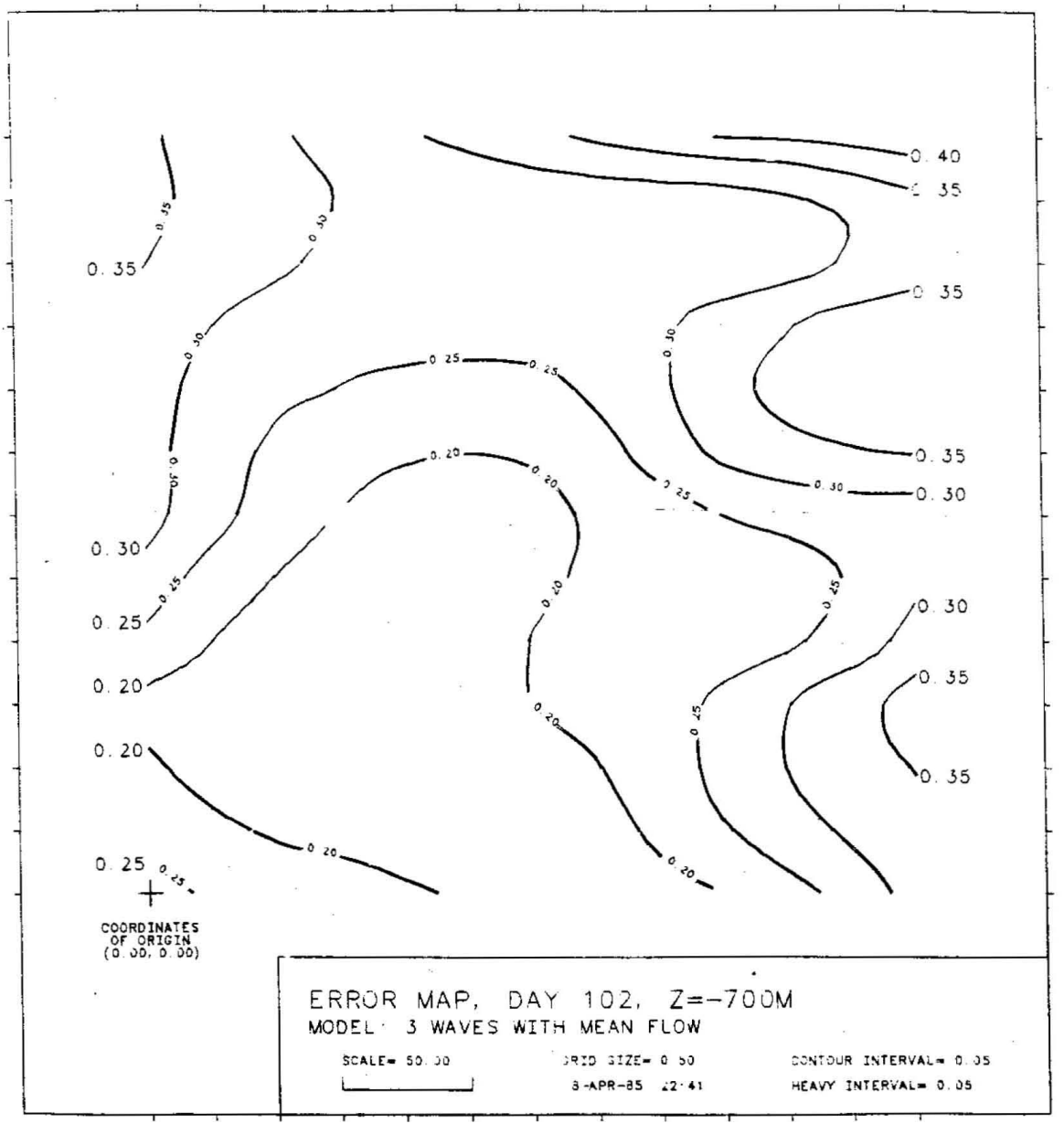

Figure 5.18. Error map, showing contours of the standard deviation at a depth of 700 m of the optimally estimated sound-speed perturbations in the wave field in the experimental square on yearday 102. Contour interval is 0.05 m/s.

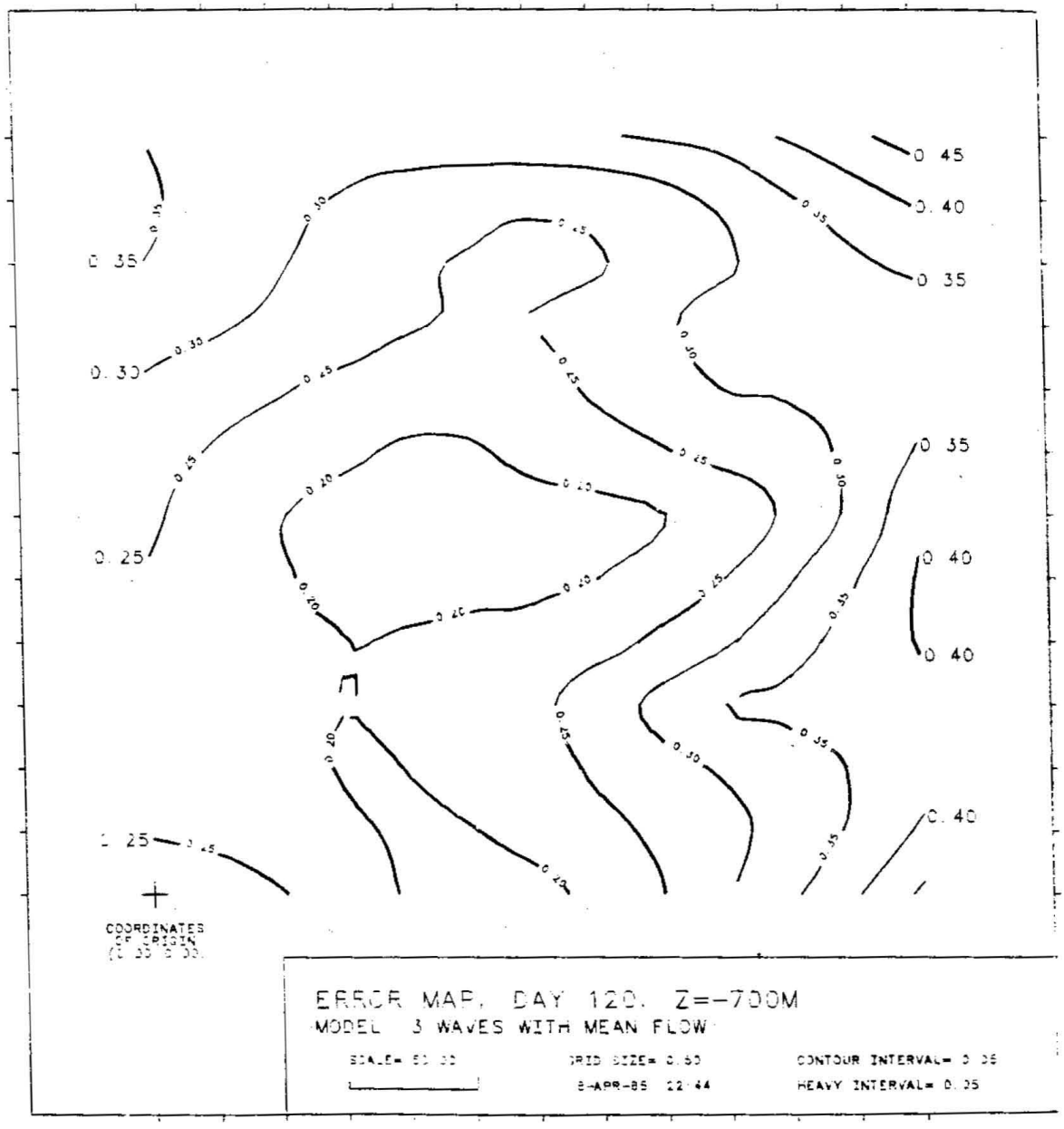

Figure 5.19. Error map, showing contours of the standard deviation at a depth of 700 m of the optimally estimated sound-speed perturbations in the wave field in the experimental square on yearday 120. Contour interval is 0.05 m/s.

CHAPTER 6

ESTIMATION OF WAVE PARAMETERS AND WAVE DYNAMICS (2): DISCUSSION AND CONCLUSIONS

6.1 Summary Of The Wave Fits

Using estimation theory and optimization techniques, we have studied the existence and dynamics of dispersive baroclinic planetary waves. The estimations were based on the profile, point and integral measurements of sound-speed (or temperature) perturbations obtained in the 1981 Ocean Tomography Experiment. Maximum Likelihood estimators that correspond to least-square fitting were employed. Many other commonly used estimation or inversion methods are analogous to the Maximum Likelihood method and the technique of least-squares, that is the generalized estimation or inversion procedure is the minimization of an objective function of a weighted sum of products of residuals as discussed in Ch. 4.

A range of one to five waves that propagate according to three plausible models were fitted to the data. The properties of the different wave fits were then compared so that the most consistent propagation model could be identified and the optimal number of existing waves could be estimated. The data set used in the fittings was derived from the measurements through filtering and data reduction.

The 'best' fit can unambigously be identified to correspond to three waves that evolved under the presence of a mean flow. The evidence of the existence of the waves is supported by the high correlation between the fit and the observed signal ($\geq$0.88) and the large amount of signal energy resolved ($\geq$78 percent), in each of the three independent data subsets. Furthermore, the high correlations and resolution cannot be a result of ill-conditioning in the system of model equations because the optimal solution for the wave parameters is unique and well-determined. As indicated in Table 5.1, the rms errors are only about 10 percent of the estimate.

6.2 Comments On The Wave Dynamics

Westward phase propagation is known to be typical of mesoscale perturbations at mid-latitudes from previous experiments. Consistently, as indicated in Table 5.1, the phases of the observed waves were all propagating westward. The corresponding group-velocity vectors have westward directions also, implying that the waves were generated somewhere to the east of the experimental region, therefore, the possibility that they were radiated by the intense Gulf Stream can be ruled out. The three baroclinic waves do not form a resonant triad since the sum or difference of the phases of two of the waves does not equal the phase of the other wave. However, the propagation of resonant baroclinic waves is still possible because they could be generated by interacting barotropic waves. The fastest oscillation that could be forced by the observed baroclinic waves would result from the interaction between the 1st and the 3rd waves and would have a period of $(1/117+1/121)^{-1}\sim 60$ days. But, since the secondary perturbation which we have observed from the moored time records of temperature has a period of 20 days (see Fig. 5.3 to 5.9), it must be due to the interaction of barotropic waves that have much higher frequency cutoffs.

In the absence of a mean flow, the short-period cutoff of first-mode baroclinic waves is approximately 160 days, e.g. (2.52), so that the waves cannot account for the high frequency content (i.e. periods of 117 and 121 days) of the data. This is well

demonstrated by the wave fits of Model 0. Although the mean current, as estimated, is very weak (approximately 2 cm/s), it strongly alters the space and time behavior of the wave-induced perturbations by producing large changes in the wave periods or frequencies (the Doppler effects have reduced the wave periods of the 3 waves by 202, 164 and 77 days, respectively). Thus, the weak mean current has played an important role on the wave propagation in the region.

The approximate solution for linear dispersive planetary waves is obtained by neglecting the nonlinear and linear-coupling terms in the horizontal-structure equations (2.43) for mesoscale motions. Let us first comment on the linearization and then discuss the linear coupling in the context of instability theory. Qualitatively, the linearization is valid when the ratio of the particle to phase speed of the waves is small when compared to unity. As the ratio decreases, so do the nonlinear effects. Therefore, by shortening the wave periods and hence increasing the phase velocities, a westward mean current can weaken the nonlinear interactions between the dispersive waves, thus making the linear approximation better. The magnitudes of the phase and particle velocities of the observed dispersive primary waves were computed and the results are presented in Table 6.1a. Furthermore, the magnitudes of the phase velocities of the waves, computed as if the mean current were absent, are also presented in the same table. It is seen that if the weak mean current were absent, the validity of

the linearization for the wave motions would be harder to justify. The pressure amplitudes of the secondary waves forced by the observed primary waves, computed using (2.66), and the pressure amplitudes of the primary waves themselves are given in Table 6.1b. The ratios of the rms pressure amplitudes of the secondary to the primary waves are approximately 1/4. Thus, there could be upto a 25 percent error in the linearized wave solution.

Table 6.1a

Magnitudes of the phase and particle velocities of the primary dispersive waves; the phase speeds in parentheses were computed by setting the mean current to zero.

wave i.d. no. i	phase speed (cm/s)	wave-induced current (cm/s)
1	6.1 (2.2)	2.2
2	3.4 (2.2)	4.0
3	18.3 (10.5)	1.8

Table 6.1b

Pressure Amplitudes Of The Primary And Secondary Waves

i.d. no. of interacting primary waves		amplitudes of primary waves (10^5kg/km s^2)		amplitudes of forced waves (10^5kg/km s^2)		ratio of rms amplitudes
i	j					
1	2	.598	1.236	.035	.014	.02
1	3	.598	.935	.023	.330	.30
2	3	1.236	.935	.044	.155	.10

Since we have observed a horizontally stratified flow with vertical shear, we shall investigate the stabilitity of the flow in the presence of wave disturbances. The corresponding instability phenomenon is baroclinic. When it occurs, the available potential energy of the sloping-isopycnal mean state is converted to the potential and kinetic energy of the perturbations. A consequence of baroclinic instability is that the wave disturbances will grow and the tilted mean-state isopycnal surfaces will become more horizontal, that is warm fluid will rise and cold fluid will sink. Another instability phenomenon, which is not considered here, is barotropic in which the kinetic energy of the mean flow is converted to the kinetic energy of the perturbations. Barotropic instability can only occur if the mean flow has a horizontal shear. The interested reader is referred to Pedlosky (1979) and LeBlond and Mysak (1978) for discussion on both barotropic and baroclinic instabilities.

Mathematically, the linear couplings in (2.43) give rise to baroclinic instability. Assuming the ocean bottom is flat, dropping the nonlinear terms, and performing a triple Fourier transformation, (2.43) can be cast as an eigenvalue problem in matrix algebra:

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} \phi_0 \\ \phi_1 \end{bmatrix} = \sigma \begin{bmatrix} \phi_0 \\ \phi_1 \end{bmatrix} \qquad (6.1a)$$

with

$$a_{11} = \frac{-\beta k}{k^2+l^2} + (u_0 k + v_0 l), \tag{6.1b}$$

$$a_{22} = \frac{-\beta k}{k^2+l^2+\lambda_1} + (u_0 k + v_0 l) + \varepsilon_{111}(u_1 k + v_1 l)\frac{k^2+l^2}{k^2+l^2+\lambda_1} \tag{6.1c}$$

$$a_{12} = u_1 k + v_1 l \tag{6.1d}$$

and

$$a_{21} = (u_1 k + v_1 l)\frac{k^2+l^2-\lambda_1}{k+l+\lambda_1} \tag{6.1e}$$

where $\varepsilon_{111}=1.932$ is evaluated by (2.36b), $\lambda_1=5.149\times10^{-4}$ km^{-2} is the inverse of the internal radius of deformation of the 1st mode squared and $\phi_0(k,l,\sigma)$ and $\phi_1(k,l,\sigma)$ are the spectra of the modal-amplitude functions of the 1st and 2nd mode perturbation pressures, respectively, as defined in (2.47). The modal-amplitude vectors of the barotropic and baroclinic mean currents are denoted by (u_0,v_0) and (u_1,v_1), respectively. For a given wavenumber vector (k,l), the wavefrequencies σ, that is the eigenvalues, are given by

$$\sigma_{\pm}(k,l)=(a_{11}+a_{22})/2 \pm [(a_{11}-a_{22})^2+4a_{21}a_{12}]^{1/2}/2 \tag{6.2}$$

Note that the coupling is caused by the baroclinic mean current only, and when coupling is neglected σ_- and σ_+ are the same as the frequencies of the dispersive barotropic and baroclinic waves, respectively. For disturbances with (k,l)'s satisfying

$$(a_{11}-a_{22})^2 < -4a_{12}a_{21}, \tag{6.3}$$

the wavefrequencies are complex. Under this condition, since σ_+ and σ_- are complex conjugates, one of them must have a positive imaginary part that corresponds to instability.

To investigate whether the observed waves are unstable, we solved (6.3) for the region of instability in the wavenumber domain, i.e., k-l plane, using the estimated values of (u_0,v_0) and (u_1,v_1). The shaded area on the k-l plane as displayed in Fig. 6.1 is the region of instability. In the figure, we also plot the locations of the observed wave disturbances. The disturbances are all located outside the shaded area, implying that the waves are stable, at least in the general area where the experiment was conducted. However, we must warn that, as the waves approach the western boundary, they may encounter changes in the direction and intensity of the mean flow such that some or all of the three waves

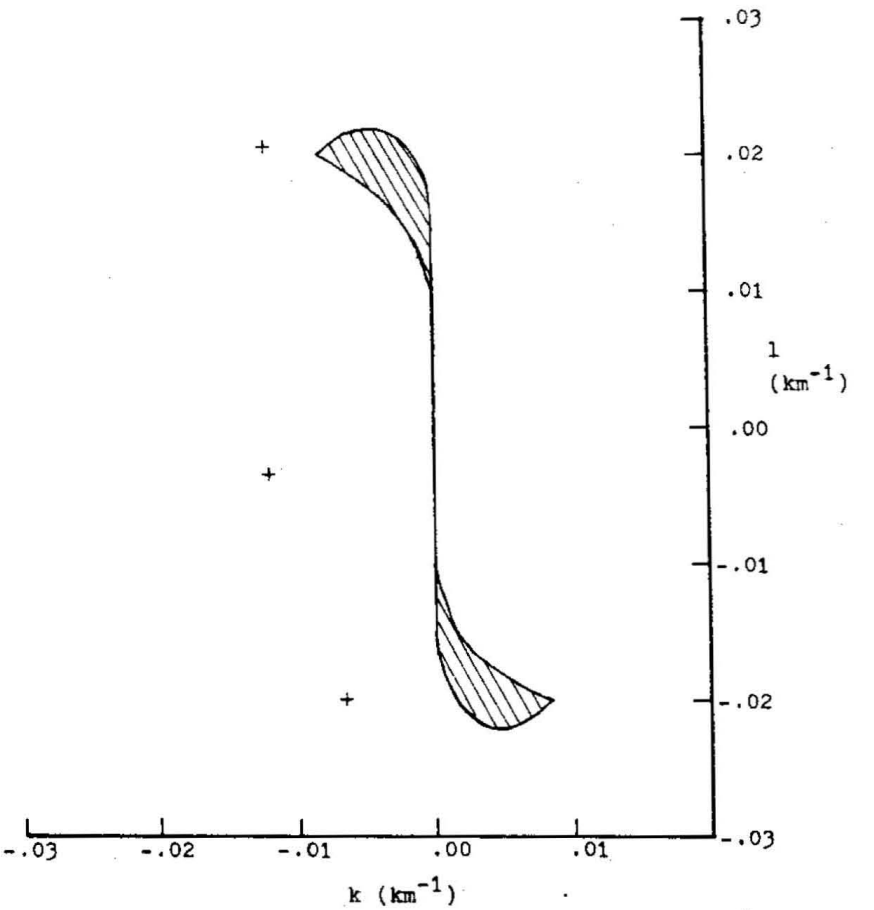

Figure 6.1. The stability of the vertically sheared mean flow in the tomographic region in the presence of wave disturbances in the first baroclinic mode. The region of instability on the (k,l)-plane, i.e. in the wavenumber domain, is the shaded area, and the wavenumber vectors of the observed disturbances (+) are in the stable region.

can become unstable and develop into intense eddies. This is because, as the mean current becomes stronger, the region of instability becomes larger; also as the flow direction changes, so does the location of the unstable region.

In spite of the fact no inconsistency between the theory and observations has been found, we recognize that a complete investigation of the wave dynamics was disallowed by the limitations imposed by the data. First, we were unable to observe any weakly nonlinear phenomenon of the baroclinic perturbations because the data occupy a time interval which is less than one wave period. Secondly, due to the insufficiency of explicit current measurements, we were unable to observe the barotropic waves.

6.3 Comparison With The MODE Wave Fits

The Mid-Ocean Dynamics Experiments, MODE-0 and MODE-1, were designed to investigate mesoscale motions and their role in the general circulation in a ~400 km square region at 28°N, 69°40'W, just north of the tomographic region. MODE-0 was a collection of several pilot studies that were carried out between 1971 and 1972 to identify the space and time scales of the energies. It was then followed by MODE-1 in the spring of 1973, which is probably the most comprehensive large-scale experiment to date. MODE-1 lasted for about 4 months.

McWilliams and Flierl (1976) have fitted the planetary-wave model to the MODE-0 and MODE-1 data sets, separately. While the former contained only current-meter records from 7 separate moorings and mostly from beneath the main thermocline, the latter was a much larger and more uniform data set, obtained from a variety of instruments: current meters, moored temperature sensors, CTD's and STD's, and SOFAR floats. The MODE-0 and MODE-1 data sets have durations of 3 and 4 months, respectively. In constrast to the observational system deployed in the tomographic experiment, the MODE arrays consisted of spot measurements only, which unlike the acoustic travel-time measurements, could be severely contaminated by undesirable small-scale features.

In the same way as our study but for weighting, McWilliams and Flierl chose the optimal wave parameters to minimize a quadratic

error norm for the differences between the data and the fit. Instead of specifying the weighting factors in the error-norm minimization according to the noise variances, as was done in our stochastic estimates, they have assigned equal weighting to each datum of the same type and made total data energies equal for all types when incorporating data of different types. Under the circumstances, we believe that their estimates do not differ significantly from the stochastic maximum-likelihood estimates, because estimates are, in general, not sensitive to the choice of weighting factors when the number of data is much larger than the number of unknown parameters.

The best MODE-0 fit has a high correlation of ~0.8 with the data and accounted for over half (~60 percent) of the data energy. It consisted of a pair of barotropic waves and no baroclinic waves, propagating in the absence of mean flow. The reason for not being able to observe any baroclinic waves is probably that MODE-0 was primarily an experiment of the lower layer (below the main thermocline) where the barotropic-mode kinetic energy dominates. Although a few current-meter records from the main thermocline were available, they came from only two horizontal locations. Therefore, they were not adequate for resolving baroclinic waves, since each wave involves at least 4 free parameters. In contrast, the tomographic experiment was primarily for the observations of the baroclinic modes. In the experiment, the acoustic array, the CTD casts and the moored temperature sensors and recorders, all probed the temperature field in which the baroclinic-mode effects dominate,

and for the same reasons as above, the current data from only 2 horizontal locations were inadequate for resolving barotropic waves.

Nonlinear interactions within the MODE-0 wave fit and our wave fit to the data of the tomographic experiment were found to be weak: forced wave amplitudes were predicted by the weakly nonlinear theory to be about 20 percent of the primary wave amplitudes. Thus both sets of waves represent fully consistent linear solutions.

On the other hand, both barotropic and baroclinic waves were observable by the MODE-1 array that contained both adequate current and temperature measurements. The best MODE-1 fit, having a correlation of ~0.7 and accounted for 1/2 of the data energy, has a pair of barotropic and a pair of first-baroclinic waves. Consistent with the MODE-0 fit, no significant energy of the mean flow was found. However, unlike the other two fits, nonlinear interactions were found to be of marginal but uncertain importance within the MODE-1 fit: forced wave amplitudes were predicted to be large and comparable to the primary wave amplitudes. But, by searching in the data for the forced waves with the given frequencies and wavenumbers, McWilliam and Flierl have found no significant energy in them. To explain this, McWilliams and Flierl suggested that the nonlinear transfers of energy might have acted in such a way as to preserve crucial features of the linear solution, empirically.

From the results of the 3 wave fits, we can summarize the dynamics of the mesoscale motion in the general area where MODE-0, MODE-1 and the tomographic experiment were conducted as follow:

(1) The motion appears to be dominantly wave-like: planetary waves have consistently accounted for more than and about 1/2 of the total signal energies observed in different places and during different time periods.

(2) The vertical structure is dominated by the barotropic and the first baroclinic modes, with the latter containing the greatest fraction of the available potential energy among all the vertical modes.

(3) Locally, the space-time behavior of the motion is well predicted by the dispersion relation, i.e. linear dynamics. But, as the lengths in space and time considered increase, the linear prediction becomes less accurate; this is demonstrated by the fact that the MODE wave fits, which involved a larger region and longer durations, have poorer quality (i.e. smaller correlations and less signals accounted for) than our wave fit. Thus, planetary wave propagation is strictly a local phenomenon.

(4) Most of the waves observed in the three experiments have westward group velocities, suggesting that wave disturbances are originated in the east.

(5) The phase propagation is generally westward, and the wave lengths of the propagating baroclinic waves are typically of order a few hundreds of kilometers.

(6) Evidence exists for the existence of a westward mean flow with diminishing flow energy towards the north: a weak westward mean flow with vertical shear was found in the tomographic region and

vanishing mean-flow intensity was found in the MODE region.

(7) In each of the three experiments, MODE-0, MODE-1 and 1981 Ocean Tomography, the data exhibited more high frequency variability than the wave fits. Therefore, nonlinear wave-wave interactions must be consequences of wave propagation.

(8) Stronger nonlinear wave-wave interactions should occur in the north, because the westward mean flow can reduce the interactions in the south by increasing the westward phase velocities there.

6.4 Comparison Of The Different Mapping Methods

Previously, Cornuelle (1983) and Cornuelle et al. (1985) have used the same acoustic and hydrographic data to map the ocean. Their mapping, however, was performed on an "objective" and "daily" basis, and the two sets of data were used separately in independent linear inversions. The mapping performed by us differs from that of Cornuelle et al. in three major aspects: (1) we have incorporated the hydrographic and the moored temperature data together with the acoustic data in the same inversions, (2) this mapping is "subjective" and takes into account the time-dependence of the field, and (3) the system being solved here is nonlinear with respect to the unknown parameters. By "subjective" mapping, as oppose to "objective" mapping, we mean that the space-time relation imposed on the unknown field in the inversion of data is a deterministic one.

In this section, we will first describe the method of Cornuelle et al. and discuss the differences and similarities between our method and theirs. We will then present some possible extensions of their method to take into account the time-dependence of the field. The advantages and disadvantages of the different methods will also be discussed. A discussion on the improvement on the inversion result due to the incorporation of the spot measurements will be presented in the next section. For the purpose of making the algebra as simple as possible in this discussion but without loss of

generality, let us assume that the positions of the acoustic moorings are accurately known in the following mathematical formulations. (A discussion on the effect of unknown mooring motions on the estimate of δc is presented separately in Ch. 7.)

Cornuelle et al. wanted to obtained the best possible estimate of the perturbation field $\delta c(\underline{x},t_k)$ of sound speed in space $\underline{x}=(x,y,z)$ on the days $t=t_k$'s of the acoustic transmissions, based on the acoutic data alone. They have chosen a linear estimator and defined the best estimate to have minimum variance. Their method of inversion is analogous to the objective mapping of Bretherton et al. (1976), in which a specification of the autocorrelation function of the unknown field is required. Cornuelle et al. have assumed that the unknown field $\delta c(\underline{x},t)$ to be horizontally homogeneous and temporally uncorrelated. Based upon the analysis of Richman et al. (1977) on the MODE-array data, they have taken the horizontal autocorrelation function to be Gaussian in shape with a decay scale of 100 km. Vertically, they have chosen to represent δc by the empirical orthogonal modes derived from the MODE-hydrographic data. Thus, the correlation function can be expressed as

$$<\delta c_i(x,y,t)\delta c_i(x',y',t')>=\sigma_i^2\delta(t-t')e^{-[(x-x')^2+(y-y')^2]/(100\ \mathrm{km})^2}\ ; \quad i=1,2,3,\ldots, \qquad (6.4)$$

where δc_i represents the horizontal structure of the sound-speed perturbation associated with the ith mode and σ_i^2 is the

expected energy of δc_i.

The tomographic system solved by Cornuelle et al. is linear and may be cast parametrically, at time t_k, as

$$\underline{\delta t}^o(t_k) = \underline{A}\,\underline{a}(t_k) + \underline{v}(t_k) \tag{6.5}$$

where $\underline{\delta t}^o(t_k)$ is an mx1 data vector containing the observed travel-time perturbations, $\underline{v}(t_k)$ represents the noise vector, and $\underline{a}(t_k)$ is an nx1 parameter vector to be estimated, containing the unknown amplitudes of the sinusoidal wavenumber components of δc_i. Unlike the other quantities in (6.5), the linear operator $\underline{A}$, that is an mxn weighting matrix, is time-independent, and $\underline{A}$ can be evaluated using (3.8). Because δc_i's are spatially homogeneous, the Fourier components in the wavenumber spectra are uncorrelated, implying that the time-independent covariance $\underline{C}_{\underline{a}}$ of $\underline{a}(t)$ is a diagonal matrix. Clearly, an advantage of choosing to estimate $\underline{a}$ instead of the δc_i's themselves is the minimization of the storage area required in the computer. Since the system (6.5) is linear and the sound-speed perturbation and noise are uncorrelated, the Gauss-Markov Theorem immediately asserts that among all linear estimates, the one with the smallest variance is

$$\underline{a}^*(t_k) = \underline{C}_{\underline{\Delta a}^*}(t_k)\underline{A}^T\underline{C}_{\underline{v}}^{-1}(t_k)\underline{\delta t}^o(t_k) \tag{6.6}$$

where

$$\underline{\underline{C}}_{\Delta a}{}^{*}(t_k) = [\underline{\underline{A}}^T \underline{\underline{C}}_v^{-1}(t_k)\underline{\underline{A}} + \underline{\underline{C}}_a^{-1}]^{-1} \qquad (6.7)$$

is the error-covariance matrix of $\underline{a}^*(t_k)$ and $\underline{\underline{C}}_v(t_k)$ is the covariance matrix of the noise $\underline{v}(t_k)$ (Liebelt, 1967). An interesting fact is that the same estimate can be obtained by maximizing (minimizing) the corresponding likelihood (objective) function. This is not surprising, however, because as we may recall from the discussion in Ch. 4, when the system is linear and the a priori information is incorporated as data in the system, the maximum-likelihood estimate has the lowest theoretically attainable variance. Therefore, an obvious similarity between the method of Cornuelle et al. and ours is that they both compute maximum-likelihood estimates. However, they did not consider the time-dependence of the field in their inversions; their estimates thus were three-dimensional ones.

The generation of a four-dimensional estimate is more desirable. One reason is that the quality of the estimate of the unknown field is generally improved when the set of observations used in the inversion is enlarged. In the detection of narrow-band planetary waves from the data, we have mapped the ocean on a subjective and four-dimensional basis, by imposing that the local sound-speed field is predominantly perturbed by the waves. That is, in the inversions, we have required the wavenumber spectra to be sharply peaked at some wavenumbers and the spectra at different times to be related by the dispersion relationship. It is

understood from inverse theory that, if the unknown function is an impulse, the best linear estimate of the function will generally contain side lobes in addition to a main lobe at the location of the impulse (Wiggins, 1972, Wunsch, 1978, etc.). The leakage of energy to the side lopes and the broadening of the main peak is a consequence of the lack of determining power which is always associated with an underdetermined system. The implication is that narrow-band planetary waves cannot be adequately resolved by directly estimating the parameter vector $\underline{a}(t)$ that represents the continuous spectral-amplitude functions. One way to eliminate the side lobes and sharpen the main lobe is to reparameterize the wavenumber spectra by the location, amplitude and phase of the peaks, and this is exactly what we have done to implement the narrow-band constraint in the inversions.

The narrow-band constraint transforms the underdetermined linear systems at different t_k into one overdetermined, nonlinear system. The linearization of the nonlinear system with respect to the unknown wavenumbers is not valid because the phase functions of the waves can be of order one or bigger at large distances and times, implying that we cannot use standard direct techniques such as Gaussian elimination and the singular-value decomposition, and must resort to the use of iterative minimization methods for the inversions. We prefer gradient methods over other iterative methods because they guarantee convergence (Ch. 4).

The error covariance of the estimate associated with the linear system (6.5) does not dependent on the data and the estimate itself, but only on the statistics of the unknown field and noise, and the geometry of the acoustic array. Difficulty in the analysis of variance increases once the estimation problem becomes nonlinear. In fact, the variance of our nonlinear wave fit could not be obtained before the estimate was computed. Therefore, in the design of tomographic experiments, it is definitely more convenient to work with the linear systems. However, the wave fit accounts for the dynamics.

While Cornuelle et al. have adopted the empirical modes (derived from the MODE data) as the vertical basis of the sound-speed perturbations, we have, instead, adopted the analytical modes of Rossby waves. An advantage in using the analytical modes is that the corresponding horizontal-structure equation can readily be obtained from the literature. The 1st and 3rd empirical modes strongly resemble the 1st and 3rd baroclinic analytical modes, and the 2nd empirical mode is strongly surface-intensified. (The empirical modes are ordered according to the ratios of their potential energy, with the most energetic one being defined as the 1st mode.) The first four empirical modes were used by Cornuelle et al. (1985) and assigned equal energy a priori; however, their inversions have yielded a result of 1 : 0.1 : 0.05 for the ratios of the energy of the first three modes, showing consistently that the vertical structure is dominated by the 1st baroclinic mode.

Moreover, the amplitudes of the higher modes were poorly determined, because most of the ray paths identified did not penetrate into the mixed layer to sense the surface-intensified mode, and the other higher modes are basically transparent to acoustic tomography (see Ch. 3 for the discussion). Our inversions, therefore, have not attained a poorer vertical resolution although only the 1st baroclinic analytical mode was used.

In objective mapping, the experimental noise basically consists of the measurement and internal-wave related errors that generally have zero expected values. However, in subjective mapping, the additional error introduced by the idealizations and assumptions used in building the dynamical model may have a nonzero statistical average. A consequence of the zero-mean hypothesis on the errors that in reality have nonzero expected values is the generation of bias error in the estimate. To illustrate this, let us suppose that the model equations $\underline{\delta t}^o = \underline{f}(\underline{p}) + \underline{v}'$ associated with a pure acoustic detection of narrow-band planetary waves can be linearized about the true values $\underline{p}_t$ of the wave parameters $\underline{p}$, so that the expectations of $\underline{\delta t}^o$, $\underline{p}$ and $\underline{v}'$ are related by, approximately,

$$\langle \underline{\delta t}^o \rangle = \frac{\partial \underline{f}(\underline{p}_t)}{\partial \underline{p}} \quad \underline{p}_t + \langle \underline{v}' \rangle \ . \tag{6.8}$$

After solving the linearized system and then using (6.8), the expectation of the maximum-likelihood estimate $\underline{p}^*$ can be written

approximately as

$$\langle \underline{p}^* \rangle = \underline{p}_t + \underline{b} \tag{6.9a}$$

where

$$\underline{b} = [(\frac{\partial \underline{f}(\underline{p}_t)}{\partial \underline{p}})^T \underline{C}_{\underline{v}'}^{-1} (\frac{\partial \underline{f}(\underline{p}_t)}{\partial \underline{p}})]^{-1} [\frac{\partial \underline{f}(\underline{p}_t)}{\partial \underline{p}}]^T \underline{C}_{\underline{v}'}^{-1} \langle \underline{v}' \rangle \tag{6.9b}$$

is the bias of the estimate and $\underline{C}_{\underline{v}'}$ is the covariance of $\underline{v}'$. Clearly, the bias exits when $\langle \underline{v}' \rangle$ is not zero.

In spite of the generation of bias in the estimate, subjective mapping has its appeal. By trying many different dynamical models in the inversions, the data can make diagnoses for plausible dynamics. Hence, one can learn the dynamics of the field directly from the inversions and then use the knowledge gained to make model corrections. In fact, the generation of bias is not of major concern, since when the model used is accurate, the bias will be small. Moreover, the estimate generated by objective mapping is also biased. Using (6.5), (6.6) and (6.7), we can easily show that the expectation of the objective estimate of $\underline{a}(t_k)$ is given by

$$\langle \underline{a}^*(t_k) \rangle = [\underline{A}^T \underline{C}_{\underline{v}}^{-1}(t_k) \underline{A} + \underline{C}_{\underline{a}}^{-1}]^{-1} \underline{A}^T \underline{C}_{\underline{v}}^{-1}(t_k) \underline{A}\ \underline{a}_t(t_k) \tag{6.10}$$

where $\underline{a}_t(t_k)$ is the true value of $\underline{a}(t_k)$. Clearly, the

objective estimate is biased, i.e. $\langle \underline{a}^*(t_k) \rangle \neq \underline{a}_t(t_k)$, unless no a priori information is asserted, that is unless the a priori covariance $\underline{C}_{\underline{a}}$ approaches infinity. But, if $\underline{C}_{\underline{a}}$ approaches infinity, so will the error covariance $\underline{C}_{\Delta\underline{a}^*}$ of $\underline{a}^*$ (which is expressed in (6.7)), because $\underline{A}^T \underline{C}_{\underline{v}}^{-1}(t_k) \underline{A}$ is singular. As a matter of fact, sufficient a priori information must be supplied to generate enough bias to ensure the stability of the inversion.

The inversion method of Cornuelle et al., which uses the linear minimum-variance criterion for the estimates, in principle, can be modified to become four-dimensional. An objective approach is the implementation of the time correlation of the field into (6.4) and the expansion of system (6.5) to include observations at other times. Let us suppose that there are N+1 equally spaced data points in each time record of travel-time perturbation, so that the expanded system can be cast as

$$\underline{\delta t}'^{o} = \underline{A}' \underline{a}' + \underline{v}' \tag{6.11a}$$

where

$$\underline{\delta t}'^{o} = \begin{bmatrix} \underline{\delta t}^{o}(t_0) \\ \vdots \\ \underline{\delta t}^{o}(t_N) \end{bmatrix}, \quad \underline{A}' = \begin{bmatrix} \underline{A} \\ \vdots \\ \underline{A} \end{bmatrix}, \quad \underline{a}' = \begin{bmatrix} \underline{a}(t_0) \\ \vdots \\ \underline{a}(t_N) \end{bmatrix} \text{ and } \underline{v}' = \begin{bmatrix} \underline{v}(t_0) \\ \vdots \\ \underline{v}(t_N) \end{bmatrix} \tag{6.11b}$$

Once again, the linear minimum-variance estimate $\underline{a}'^*$ of $\underline{a}'$ can be found by applying the Gauss-Markov Theorem, at least in theory.

However, the implementation of the estimation procedure on available computing machinery may not be feasible, since the storage requirements for the covariance matrices $\underline{\underline{C}}_{\underline{a}'}$ of $\underline{a}'$ and $\underline{\underline{C}}_{\Delta\underline{a}'*}$ of $\underline{a}'*$ can be large and thus the computation of $\underline{a}'*$ might be too costly. To obtain $\underline{a}'*$, we need to evaluate its (error) covariance by

$$\underline{\underline{C}}_{\Delta\underline{a}'*} = (\underline{A}'^{T}\underline{\underline{C}}_{\underline{v}'}^{-1}\underline{A}'+\underline{\underline{C}}_{\underline{a}'}^{-1})^{-1}. \tag{6.12a}$$

or equivalently, as shown in Liebelt (1967), by

$$\underline{\underline{C}}_{\Delta\underline{a}'*} = \underline{\underline{C}}_{\underline{a}'}-(\underline{\underline{C}}_{\underline{a}'}\underline{A}'^{T})(\underline{A}'\underline{\underline{C}}_{\underline{a}'}\underline{A}'^{T}+\underline{\underline{C}}_{\underline{v}'})^{-1}(\underline{\underline{C}}_{\underline{a}'}\underline{A}'^{T})^{T}. \tag{6.12b}$$

Because the system is highly underdetermined, the latter formula (6.9b), which involves the computation of the inverse of a smaller matrix should be used; the inversion of this m(N+1)xm(N+1) matrix would consume the largest portion of the total computer time required to produce the estimate. Since the time required to perform a matrix-inverse operation is approximately proportional to the cube of the row (or column) dimension of the matrix (Dahlquist and Bjorck, 1974), this four-dimensional objective mapping can be very inefficient for large N.

An alternative approach, which is subjective, is to impose a deterministic relation instead of a statistical correlation between the perturbations or the wavenumber spectra at different times. In

this case, a linear, minimum-variance, four-dimensional estimate can also be obtained if the dynamical relation is linear or can be closely approximated by a linearization at all time steps, such that

$$\underline{a}(t_{k+1}) = \underline{D}_k \underline{a}(t_k); \; k=0,1,2,\ldots,N-1. \tag{6.13}$$

With the presence of the dynamical relation (6.13), the number of independent or free parameters in (6.11) is drastically reduced, and one can choose the unknown to be the initial spectral amplitudes $\underline{a}(t_0)$ or the spectral amplitudes at any other time. As a result, the covariance matrices are nolonger overly large. Furthermore, the linear, minimum-variance, subjective estimate can be computed using an accelerated algorithm for a Kalman filter that corresponds to a sequence of predictions and reestimations at each of the time steps (Gelb et al., 1974), so that an abundance of computation time can be saved. In (6.13), the $\underline{D}_k$'s are often called the transition matrices.

A derivation of the sequential-reestimation algorithm through the minimization of the corresponding objective function is presented in Appendix, and we will demonstrate the superior efficiency of this algorithm next. In Appendix, we show that, by choosing $\underline{a}(t_N)$ to be the free parameters, the optimal estimate $\underline{a}^*(t_N)$ of $\underline{a}(t_N)$ can be obtained by sequentially computing

$$\underline{a}^{*}(t_{l+1}) = \underline{H}_{l+1}^{-1}[\underline{A}^{T}\underline{C}_{\underline{v}}^{-1}(t_{l+1})\underline{\delta t}^{o}(t_{l+1})+\underline{C}_{\underline{a}}^{-1}(t_{l+1})\underline{a}^{p}(t_{l+1})] \tag{6.14a}$$

in order of increasing l, where

$$\underline{H}_{l+1}^{-1} = \underline{C}_{\underline{a}}(t_{l+1})-[\underline{C}_{\underline{a}}(t_{l+1})\underline{A}^{T}][\underline{A}\ \underline{C}_{\underline{a}}(t_{l+1})\underline{A}^{T}+\underline{C}_{\underline{v}}(t_{l+1})]^{-1}[\underline{C}_{\underline{a}}(t_{l+1})\underline{A}^{T}]^{T}, \tag{6.14b}$$

$$\underline{a}^{p}(t_{l+1}) = \underline{D}_{l}\underline{a}^{*}(t_{l}) \tag{6.14c}$$

and

$$\underline{C}_{\underline{a}}(t_{l+1}) = \underline{D}_{l}\underline{H}_{l}^{-1}\underline{D}_{l}^{T}. \tag{6.14d}$$

There are altogether N+1 applications of (6.14) in the sequence, and in each application, the computation of the inverse of an mxm matrix is involved (as indicated in (6.14b)). Hence, the total computer time required by the sequential-reestimation algorithm is proportional to $(N+1)m^3$. Thus, when compared to the four-dimensional objective mapping, subjective mapping with a linear or linearizable dynamical relation is $(N+1)^2$ times faster. For large N, the computational cost saved by the sequential-reestimation alogrithm in performing a four-dimensional mapping can be substaintial.

One would probably consider using the economical sequential-reestimation algorithm when the sound-speed perturbations are assumed to be produced by broad-band planetary waves. However, one must be aware that the applicability of the algorithm depends critically on the validity of the linearization of the dynamical relation. When the relation is nonlinear, the error introduced by the linearizations involved at each time step demands special investigation, since the error can propagate along the reestimation sequence and be amplified. Thus, the presence of a mean flow in the tomographic region can present some difficulties in the implementation of the wave dynamics into the transition matrices $\underline{D}_k$, because the dynamical relation between the wavenumber spectra at different times is nonlinear when the intensity and direction of the flow are unknown. (However, even when the linearization is invalid, one may still estimate the broad-band spectra by iterative minimization techniques.) This broad-band, subjective mapping has yet to be performed, but it should be of interest to compare the hypothesis of broad-band to the hypothesis of narrow-band wave disturbances in describing the mesoscale fluctuations in the region.

We have used the iterative gradient method of Fletcher and Powell (1963) for our nonlinear inversions, that is the wave fits. In order to obtain the gradient vector of the objective function, which is required by the method, the gradient vectors of the wave-induced travel-time perturbations must first be computed, which involves integrating the derivatives of the wave-induced sound-speed

perturbations with respect to each of the wave parameters along all the long-range ray paths used. The method, therefore, could be very inefficient if the integration operations were to be performed at each iterative step of the minimization process. To accelerate the process, we have precalculated the matrix $\underline{A}$ of the linear system (6.5) and have it stored in the computer, so that the gradients could be interpolated by a two-dimensional cubic spline whenever they were needed. Excluding the computer time required to compute $\underline{A}$, each minimization consumed 40 to 60 minutes on a VAX 11/780. We also experimented the daily (i.e. three-dimensional) objective inversions using Gaussian-elimination techniques on the VAX 11/780 and found that each of the inversions would consume approximately 5 minutes, again excluding the time required to compute $\underline{A}$. Therefore, by projection, the time required to do a broad-band, four-dimensional inversion using the sequential-reestimation algorithm involving 8 time steps (i.e. N=8) or to do a sequence of 9 daily inversions is approximately $(N+1)x5=9x5=45$ minutes. This is quite comparable to the time required to do one minimization of the objective function of the wave parameters, with the same number of data incorporated. Finally, the time required to do a time-dependent objective mapping that incorporates the same number of data is approximately, again by projection, $(N+1)^2x45=81x45=3645$ minutes, indicating that the computational burden is huge.

6.5 Pure Acoustic Estimates

The spot observations contain some pieces of information about the waves which are independent to those detected by the acoustic array. In the wave fits, the additional independent information acts to enhance the uniqueness and reduce the variance of the estimates of the wave parameters and the corresponding sound-speed perturbations.

When the spot measurements are withheld, the estimates are degraded. In Fig. 6.2 and 6.3, we show the maps of the sound-speed estimate on yearday 83 and 120 at a depth of 700 m, generated by a fit of 3 waves of Model 1 to the travel-time data alone, and therefore corresponding to the result of a time-dependent pure acoustic inversion. The two corresponding error maps are presented in Fig. 6.4 and 6.5, showing the contours of the standard deviation of the error of the sound-speed estimate. These errors are about half the size of those errors in the time-independent acoustic maps produced by Cornuelle (1983) and Cornuelle et al. (1985), but are 2 times larger than those of the optimal fit when the spot measurements are included (see Fig. 5.17 to 5.19). Furthermore, as expected, the error maps indicate that away from the central region of the experimental area, where the ray-path density is low, the mapping ability by the acoustics diminishes. Notice also that the errors on the left half of the square where more ray paths have traversed are slightly smaller, as a result of the presence of the

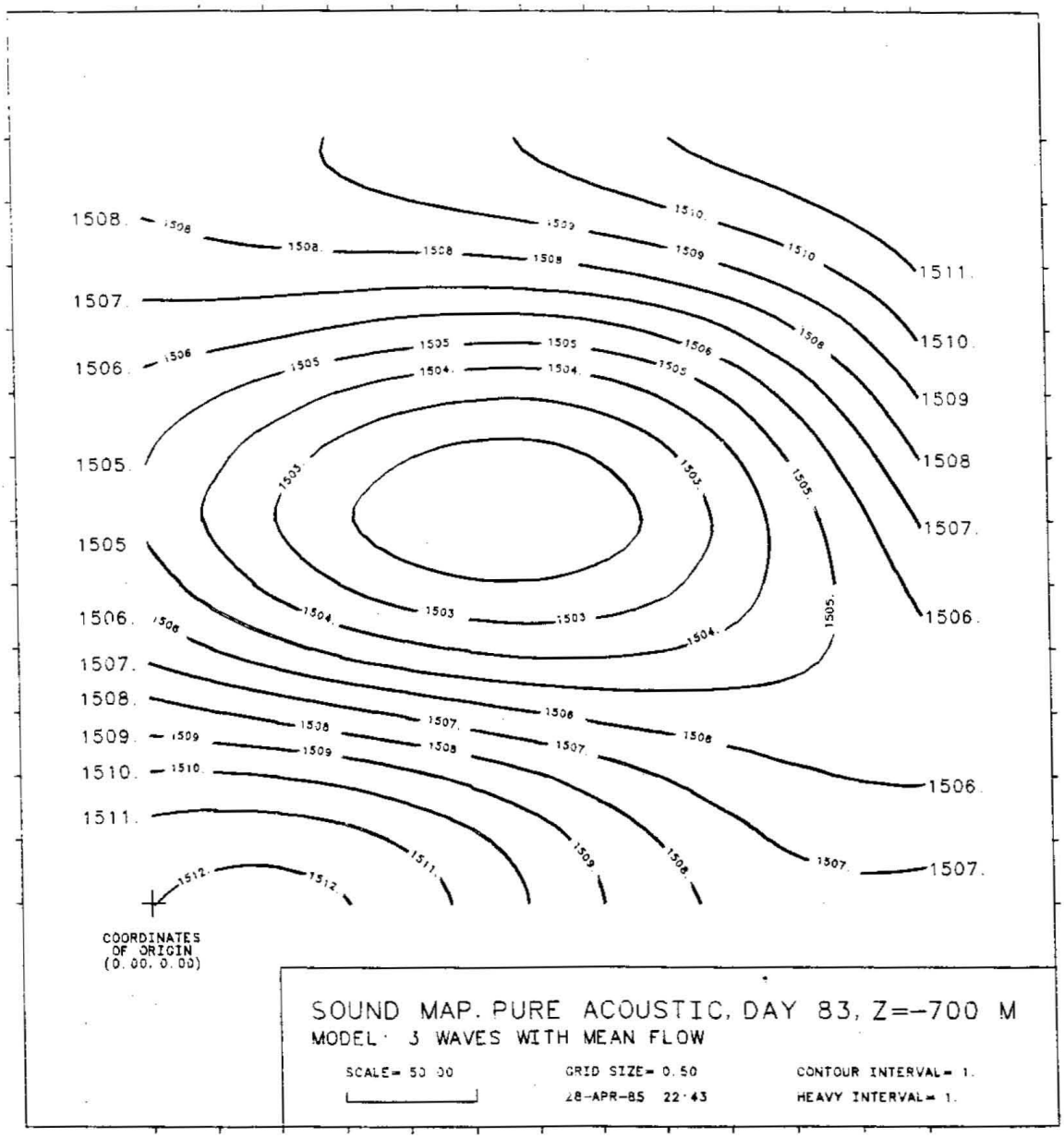

Figure 6.2. Sound-speed contours at a depth of 700 m of a pure acoustic estimate of the wave field in the experimental square on yearday 83. Contour interval is 1 m/s and the reference sound-speed at this depth is 1506 m/s.

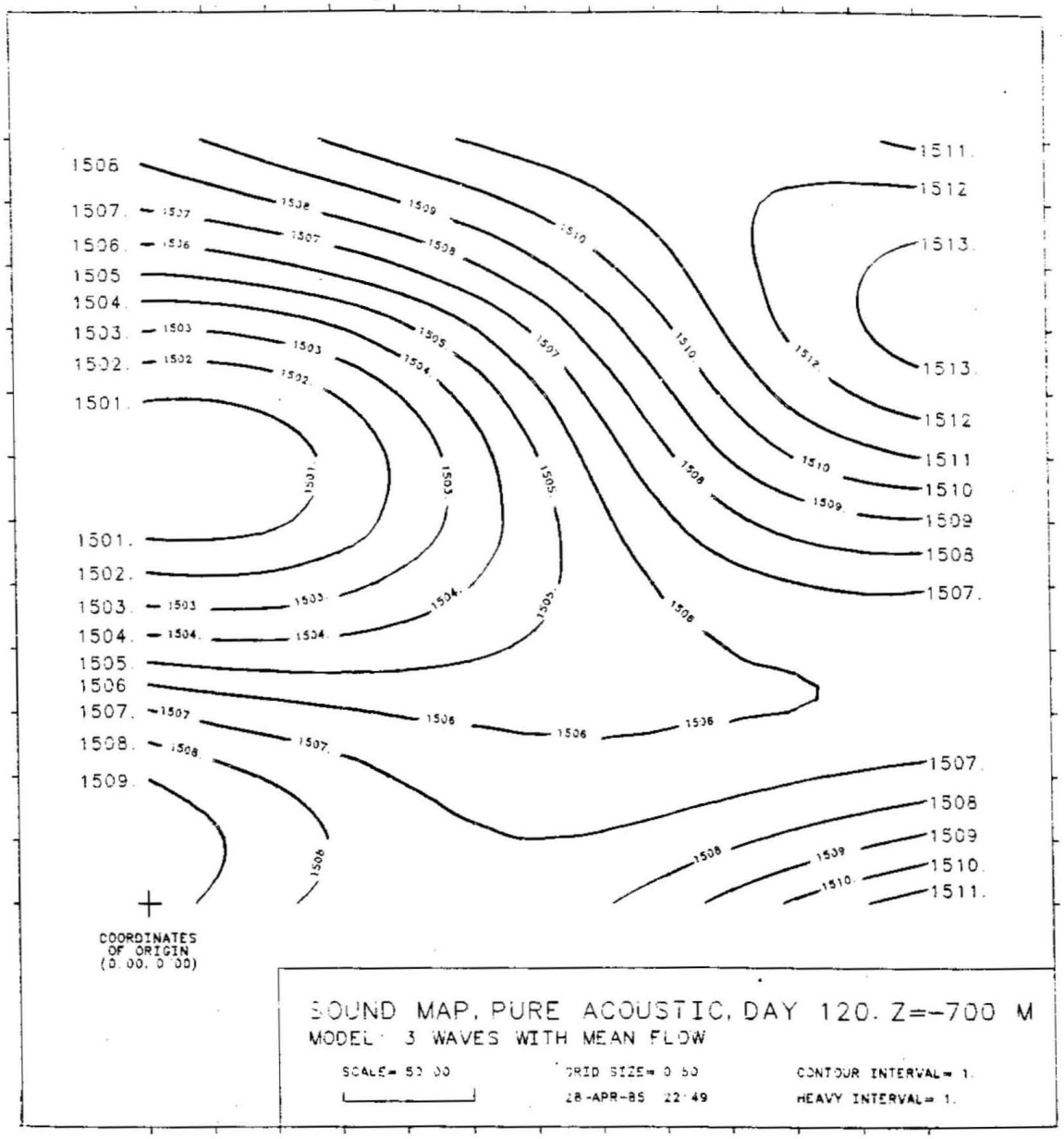

Figure 6.3. Sound-speed contours at a depth of 700 m of a pure acoustic estimate of the wave field in the experimental square on yearday 120. Contour interval is 1 m/s and the reference sound-speed at this depth is 1506 m/s.

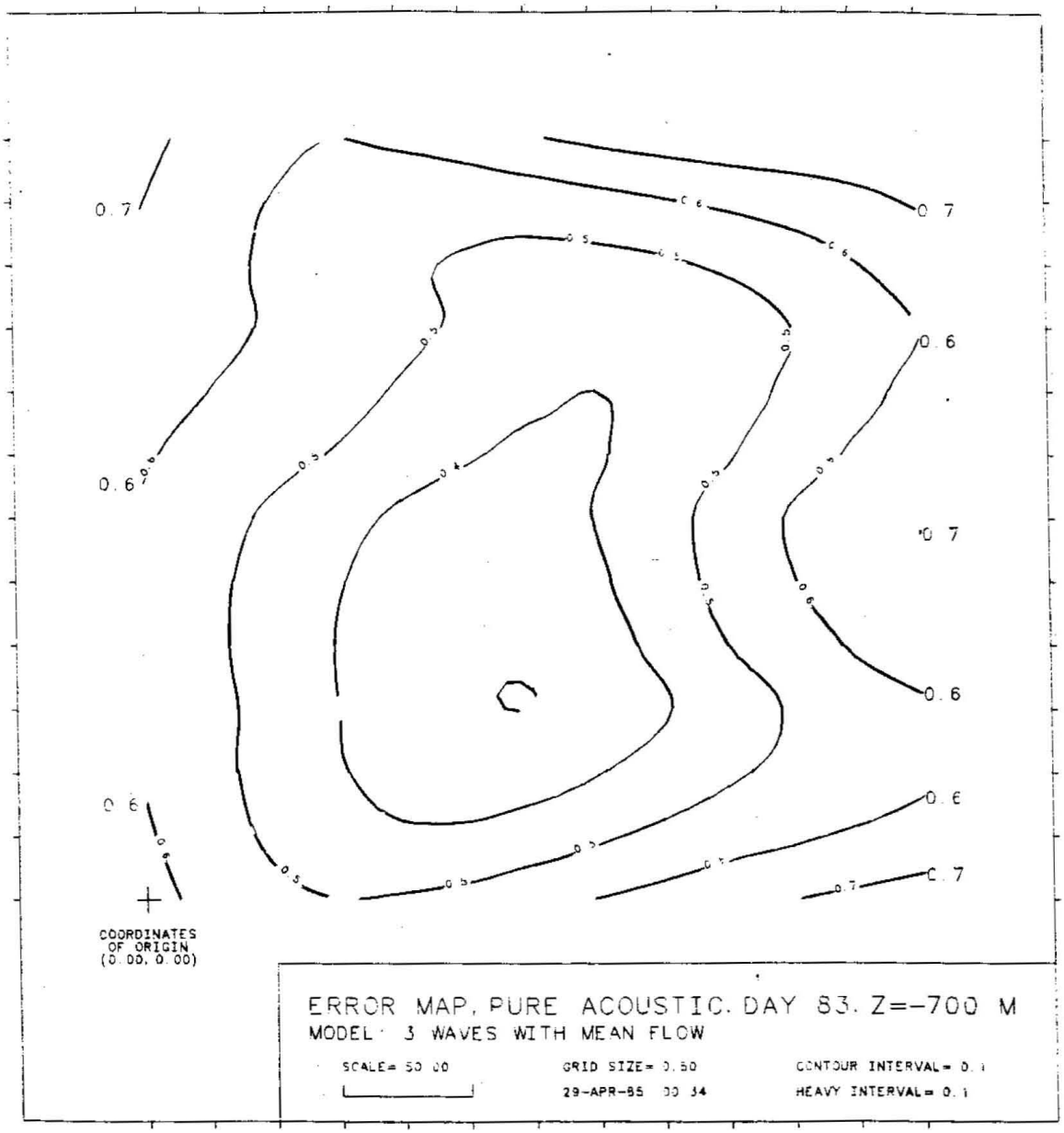

Figure 6.4. Contours of the standard deviation, at a depth of 700 m in the experimental square on yearday 83, of a pure acoustic estimate of the sound-speed perturbations in the wave field. Contour interval is 0.1 m/s.

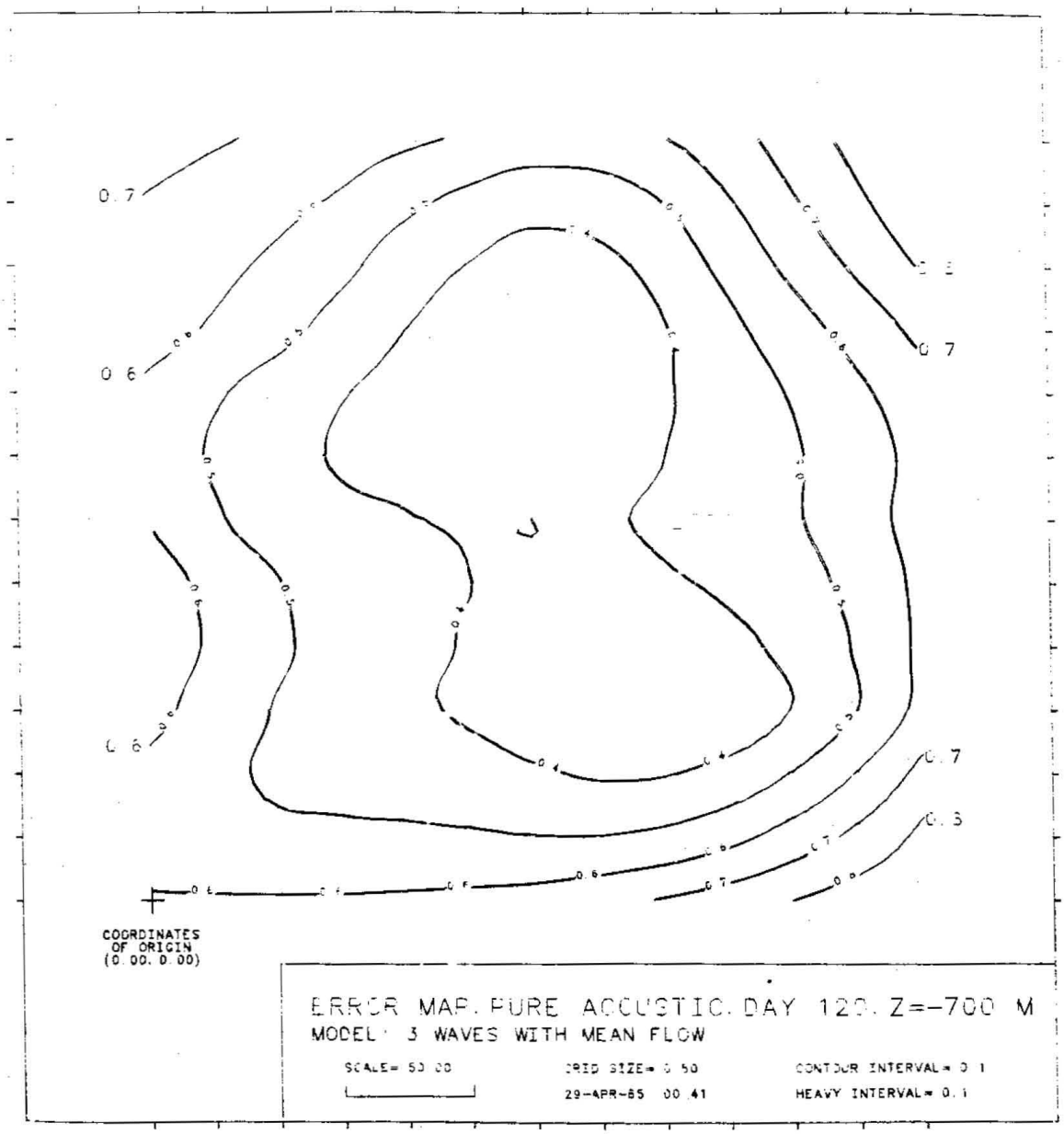

Figure 6.5. Contours of the standard deviation, at a depth of 700 m in the experimental square on yearday 120, of a pure acoustic estimate of the sound-speed perturbations in the wave field. Contour interval is 0.1 m/s.

receiver R5.

In addition to having a larger error variance, the pure acoustic estimate of the wave parameters is also nonunique. However, the different wave-parameter estimates do produce a similar pattern in the sound-speed perturbation, showing, qualitatively, an elliptical cold eddy, initially located at the center of the experimental square and slowly moving westward. Consistently, Cornuelle et al. have also observed a similar pattern from their objective acoustic maps.

Although the travel-time data obtained from this first tomographic experiment are not powerful enough to determine the wave parameters by themselves, they certainly have contributed significantly to the success of the detection of the waves. The dynamical field in the time period separating the two CTD surveys cannot be extrapolated from the surveys alone; one can hardly deduce any relation between the two CTD maps (Fig. 5.11 and 5.15) but only to observe from them that the initial cold eddy has disappeared and a front has appeared in the experimental square at the later period. Furthermore, the moored temperature time series obtained at three horizontal spots that only occupy less than 1/4 of the square cannot possibly determine the directions of wave propagation. (The fit with three waves to just the CTD and moored temperature data was found to be nonunique.) Thus, the travel-time data has provided the essential information on the westward movement of a cold pattern that links the other information.

We have learnt from simulation inversions that when the locations of the acoustic moorings are known, the wave parameters can be uniquely determined by the travel-time data alone. In the experiment, however, the acoustic moorings S4 and R5 had no mooring-motion data, and all the other acoustic moorings had some gaps in the mooring-motion data series. Therefore, the failure to track all the acoustic mooring motions has prevented the tomographic array to perform optimally in the wave observation.

New et al. (1982) and Munk and Wunsch (1982) have studied the horizontal resolution of the tomographic configuration of the 1981 experiment for a perfectly navigated array, using the Backus-Gilbert method (1967, 1968 and 1970). By considering the worst case, that is without the use of a priori information such as the temporal and spatial correlations of the field, New et al. have found a minimum average resolution length of 100 km (i.e. 1/3 of the array size). By incorporting spatial correlation, Munk and Wunsch have reported a resolution length of order 50 km. Thus, the tomographic array is potentially capable of resolving waves or isolated oceanic features of lengths as short as approximately 100 km. In order to attain the same resolution, a conventional spot-observational system would require at least a total of 36 moorings, that is a minimum of one mooring per 50 km square (a criterion from the Sampling Principle (Steiglitz, 1974, and Bendat and Pierson, 1971)). In comparison, the tomographic system that consists of only 9 moorings is therefore more economical and adequate than a conventional system for ocean

monitoring when the acoustic moorings are tracked accurately.

Besides resolution, an important measure of system performance is the variance of the estimate. For perfect navigation of the acoustic array, a first-baroclinic perturbation signal of 2 m/s (rms) at 700 m depth, a horizontal Gaussian correlation of the field with a decay scale of 100 km, no correlation in time, and a noise level of 5 ms, the standard deviation (i.e. the square root of the variance) of the pure acoustic estimate at a depth of 700 m is contoured in Fig. 7.1. It is seen that over 60 percent of the tomographic region, mostly in the middle of the square, has a standard deviation which is below ±0.4 m/s or less than 20 percent of the signal. However, the error increases to 40 percent near the western and the eastern boundaries where the arrays of sources and receivers are located. The increase in error is due to the fact that the ray-path density is the lowest near the acoustic moorings. It is obvious that the system performance can be improved efficiently by mounting temperature recorders on the acoustic moorings. In doing this, the number of moorings used in the observational system stays the same but the variance is reduced in the areas near the moorings.

6.6 Concluding Remarks

The main purpose of this study has been to investigate the existence and dynamics of planetary waves in the tomographic region, and to find out whether the waves, when present, could be detected from the data of the experiment. The detection process consisted of the estimation of wave parameters and diagnosing the plausible wave dynamics with the data. From the result of the estimation, we have come to the following conclusions: (1) stable and dispersive planetary waves did exist, at least as a local phenomenon in space and time, (2) the wave propagation was strongly affected by the local mean flow, even though the mean flow was weak (a few cm/s), and (3) due to the existence of some experimental deficiencies such as untracked mooring motions, the tomographic observational system alone was unable to detect the waves; however, the spot observations have provided the additional information needed to make the detection successful.

In this particular study, we have demonstrated the usefulness of imposing dynamical constraints in the inversions of data. That is, by imposing different but plausible dynamics, one can learn the dynamics of the field directly from the inversions. The incorporation of dynamics may happen to convert a linear system to a nonlinear one, as this was our case, but we should not be disturbed by this consequence, since there are many iterative minimization techniques available for nonlinear inversions. However, in the

design of future tomographic experiments, it is still recommended to work with linear systems whenever possible, because the corresponding sensitivity analyses are much simpler and analytically more tractable (Ch. 7 illustrates the use of linear systems for one such analysis).

A significant consequence of the incorporation of the wave dynamics was the observation of the barotropic component of the local mean flow through the dispersion relationship, which would otherwise be impossible to observe due to the lack of explicit current measurements (unless some other assumptions were made, such as the level of no motion). We have also obtained an estimate of the baroclinic component of the mean flow, corresponding to a westward shear flow of the 1st baroclinic mode. Supporting evidence for the presence of such a sheared mean flow in the tomographic region can be found in Cornuelle et al. (1985): they have computed the difference between the average sound-speed profiles in the tomographic and MODE regions, and the differenced profile strongly resembles the first baroclinic mode (Fig. 3.6); moreover, it is negative and negative perturbation implies the flow direction is westward. In Fig. 6.6, we show the profiles of the mean current obtained in the optimal wave fit.

One of our goals was to investigate whether planetary waves could be detected by acoustic tomography alone. It was, perhaps, a little bit disappointing to find out that the tomographic system deployed in the experiment was not able to do so alone, that is to

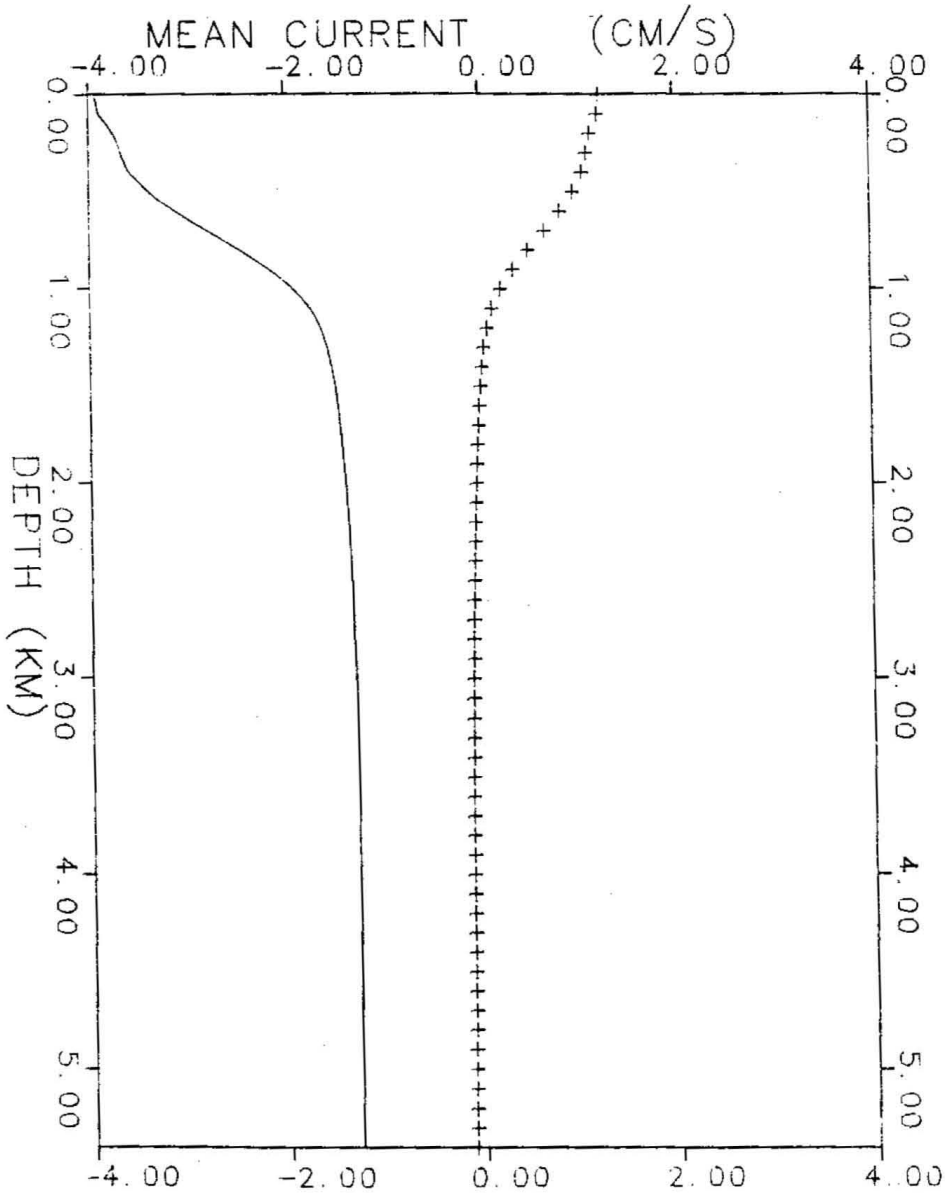

Figure 6.6. The estimated eastward (_____) and northward (+ + +) mean current profiles in the tomographic region.

determine the wave parameters uniquely. But, we must keep in mind that this was only the first field test of such observational system, and therefore the system was far from being perfect. It can be shown in computer simulations that the waves would have been detected if the noise level was reduced to ~5 ms or the mooring positions were accurately navigated, suggesting that the tomographic system is potentially capable of detecting such waves by itself.

Obviously, the spot-measurement system deployed was also unable to detect the waves alone. The reason is that the system did not obtain any information on the wave field over a long period (~40 days) between the two CTD surveys, except at three horizontal spots where the midwater temperature recorders and sensors were moored. As to the spot-measurement system, the inclusion of the acoustic data provided the missing information needed to make the detection successful. In view of the pure acoustic objective maps in Cornuelle (1983), Cornuelle et al. (1985) and the result of our pure acoustic wave fits, we may describe the acoustic data as containing the information of the westward movement of a cold pattern. This information has filled the gap between the two CTD surveys and the moored temperature data at three horizontal spots to give a unique estimate of the wave parameters.

In retrospect, the major obstacle to understanding the large-scale fluctuations in the ocean interior has been the difficulty in observing them. Traditional observational systems by

themselves are not adequate for large-scale monitoring, because an excessive amount of ship time and too many instruments would be required to attain the proper resolution of the field. The newly invented technique of acoustic remote sensing, however, holds great promise (Munk and Wunsch, 1979, and The Ocean Tomography Group, 1982). A full tomographic system is much more cost-effective than a full spot-measurement system and has the potential to provide adequate mapping by itself, as has been demonstrated by Cornuelle et al. (1984). In this study, we have further demonstrated that a tomographic observational system, when incorporated with sparse spot measurements and the plausible dynamics of the field, is certainly capable of making observations of large-scale phenomena.

CHAPTER 7

THE ERROR OF THE TOMOGRAPHIC INVERSE SOLUTION IN THE PRESENCE OF UNTRACKED MOORING MOTIONS

7.1 Introduction

In this chapter, we investigate the error of the optimal solution δc^* for the large-scale sound-speed perturbation δc in space $\underline{x}=(x,y,z)$, attained via a pure acoustical inversion based on the travel-time data $\underline{\delta t}$ measured at one moment in time. In particular, we study the error variance $\langle \Delta \delta c^{*2} \rangle = \langle (\delta c^* - \delta c)^2 \rangle$ of δc^* in the presence of untracked horizontal random motions $\underline{\delta x}$ of the moored acoustic sources and receivers. Since we do not consider time-correlated mappings of δc, the time dependence of δc, $\underline{\delta x}$ and $\underline{\delta t}$ is suppressed.

Since the observed travel-time perturbations $\underline{\delta t}$ contain information on oceanic perturbations integrated along the ray paths and since the integration automatically filters small-scale oceanic perturbations, $\underline{\delta t}$ are prominent candidates for the data to be used in estimating the large-scale sound-speed perturbations in mid-oceans. However, in using $\underline{\delta t}$, the fluctuating horizontal motions of the sources and receivers $\underline{\delta x}$ must be taken into special consideration because the dominant portion of $\underline{\delta t}$ is produced by $\underline{\delta x}$ rather than δc. While a horizontal mooring displacement of 200 m perturbs the travel time by more than 100 ms, a typical mesoscale

eddy field perturbs the travel times by only about 25 ms in a 300 km transmission. The large travel-time perturbations produced by $\underline{\delta x}$ cannot be modelled as part of the experimental noise, because this will only cause the estimate of δc to have an unacceptablely large error variance. The vertical component of mooring motion is not considered here because it is usually smaller and produces insignificant travel time perturbations.

In order to estimate $\underline{\delta x}$ and δc accurately, the use of acoustical navigational systems for tracking mooring positions was recommended by Munk and Wunsch (1979) and deployed by The Ocean Acoustic Tomography Group during the 1981 Ocean Acoustic Tomography experiment. The idea is to estimate δc based on the corrected travel time data in which the large noise induced by the mooring motion is removed. However, tracking data can be missing because of instrument failure; in that case, the best estimate of δc is found by treating the travel-time perturbations induced by $\underline{\delta x}$ also as signals in a inversion in which both δc and $\underline{\delta x}$ are estimated, simultanueously (Cornuelle, 1983, and also see Chapter 4 for the discussion on design-parameters subject to errors). In this way, an optimal estimate $\underline{\delta x}^*$ of $\underline{\delta x}$ is also found; $\underline{\delta x}^*$, with no doubt, is a reliable estimate since the corresponding signals dominate in the data. However, the objective of Ocean Acoustic Tomography is to get a reliable estimate of δc rather than $\underline{\delta x}$, and we can expect some trade-offs between the quality of the two estimates, for large $\underline{\delta x}$ can upgrade $\underline{\delta x}^*$ and degrade δc^* at the same time.

Although the simultaneous estimation of $\underline{\delta x}$ and δc is the last resort for missing tracking data, it is worthwhile and interesting in considering the economic aspects of Ocean Acoustic Tomography, to ask whether reliable mapping of δc can be generated without the deployment of navigational systems for tracking mooring motions at all. A general answer to the above question cannot be given because it depends upon particulars: the amount of available information concerning δc, such as the statistics of its horizontal and vertical structure in the ocean of interest, the smallness of the experimental noise compared to the oceanic signal, the tomographic configuration (geometrical arrangement of acoustic sources and receivers in the ocean), and the variance σ_x^2 of $\underline{\delta x}$ (which depends on the type of moorings used and the forces acting on them), all contribute to the answer. Thus, a problem in engineering design is to decide whether tracking mooring positions is necessary or not, prior to conducting an experiment in a selected ocean, with the available statistics of δc and $\underline{\delta x}$, and a selected tomographic configuration. The decision can be made only by computing $\langle \Delta \delta c^{*2} \rangle$ in numerical simulations and seeing if the error is tolerable.

The main purpose of this chapter is to show that there is an upper bound for $\langle \Delta \delta c^{*2} \rangle$ as a function of σ_x^2 and this upper error variance bound is rapidly reached with slowly increasing σ_x^2, implying that the error of δc^* is effectively independent of mooring motion once the latter has reached a critical value. This result simplifies the decision making process because only the

upper error variance bound is important for the determination of whether tracking mooring motions is needed or not, regardless of the size of σ_x. In the next section, the system of equations with unknown δc and $\underline{\delta x}$ are formulated, and the system is then used for the derivation of the analytical expressions for δc^* and its error variance in Section (7.3). Also in Section (7.3), the upper error variance bound is shown to exist. This upper error variance bound coincides or approximately coincides with the error variance of a solution for δc that is estimated with the "differenced system". The differenced system, in which $\underline{\delta x}$ is eliminated, consists of a set of "differenced model equations" that relates δc to the "differenced travel time perturbation data". In the elimination of $\underline{\delta x}$, one of the model equations associated with a resolved ray path for each of the source-receiver (S-R) pairs is used as a reference and substracted from the other equations associated with the other resolved ray paths for the same S-R pair. The differenced system, its solution and the error variance of its solution are presented in Section (7.4). In a computer simulated study presented in Section (7.5), we demonstrate that the upper error variance bound is rapidly reached. Conclusions are stated in Section (7.6).

7.2 The System With Untracked Mooring Motions

Suppose there are NS moored sources ($S_1, S_2, \ldots, S_{NS}$) and NR moored receivers ($R_1, R_2, \ldots, R_{NR}$) deployed in a typical mid-ocean tomographic experiment, and there are N resolved multipaths for each of the S-R pairs, so that at an instant in time, there are a total of m=qxN observed travel time perturbations with q=NSxNR and a total of u=2(NS+NR) unknown horizontal mooring displacement components. Let δt_{il} be the travel time perturbation observed from the ith ray path in the set of N resolved multipaths that connects the lth S-R pair; this ray path has a nominal trajectory given by $\underline{x}(s_{il})$ with s_{il} being the arc length along the path's trajectory. Let us define the lth S-R pair as the S_j-R_k pair with l=NR(j-1)+k. Also, denote the (eastward,northward) horizontal random mooring displacements of S_j and R_k as $(\delta x_s, \delta x_{s+1})$ and $(\delta x_r, \delta x_{r+1})$, respectively, with s=2j-1 and r=2NS+2k-1. It then follows from Cornuelle (1983) that the linearized model equation corresponding to the datum δt_{il} can be expressed as

$$\delta t_{il} = -\int_{\underline{x}(s_{il})} \frac{\delta c(r)}{c_0(z)^2}\, ds_{il} + a_{il}[(\delta x_s - \delta x_r)\cos\phi_l + (\delta x_{s+1} - \delta x_{r+1})\sin\phi_l] + v_{il}, \tag{7.1}$$

where a_{il} is the ray parameter (the sound slowness at the turning point) of the ray $\underline{x}(s_{il})$, v_{il} is the experimental noise in

δt_{il}, $c_0(z)$ is the mean sound speed profile that varies with depth -z, and ϕ_l is the direction of the horizontal line of transmission from S_j to R_k, measured in degrees (positive anticlockwise) with respect to the east-axis (x-axis).

Following Munk and Wunsch (1979), we discretize $\delta c(\underline{x})$ into an n-dimensional vector $\underline{\delta c}$ with the components being the sound speed perturbations averaged over small regions (boxes) of equal volume in the ocean, so that the term involving the continuous integration in (7.1) can be approximated as a weighted discrete sum:

$$-\int_{\underline{x}(s_{il})} \frac{\delta c(r)}{c_0(z)^2}\, ds_{il} \sim \underline{w}_{il}^T \underline{\delta c}, \tag{7.2}$$

with each component in the weighting vector $\underline{w}_{il}$ being minus the product of the length of the segment of s_{il} and the mean-square sound slowness in the corresponding box. After joining all the δx_i's in the vector $\underline{\delta x}$ such that

$$\underline{\delta x} = (\delta x_1, \delta x_2, \ldots, \delta x_u)^T, \tag{7.3}$$

(7.1) can be approximated, with the use of (7.2), as

$$\delta t_{il} = \underline{w}_{il}^T \underline{\delta c} + \underline{b}_{il}^T \underline{\delta x} + v_{il}, \tag{7.4}$$

where $\underline{b}_{il}$ is the weighting vector of $\underline{\delta x}$ that has only four nonzero

components: $\pm\, a_{i1}\cos\phi_1$ and $\pm\, a_{i1}\sin\phi_1$ in the corresponding columns as described by (7.1).

We can now proceed to write down the system of equations appropiate for the tomographic inversion. After segmenting the complete data vector $\underline{\delta t}$ into partial data vectors $\underline{\delta t}_i$'s and the complete noise vector $\underline{v}$ into partial noise vectors $\underline{v}_i$'s such that

$$\underline{\delta t}_i = (\delta t_{i1}, \delta t_{i2}, \ldots, \delta t_{iq})^T;\ i=1,2,\ldots,N, \tag{7.5}$$

and

$$\underline{v}_i = (v_{i1}, v_{i2}, \ldots, v_{iq})^T;\ i=1,2,\ldots,N, \tag{7.6}$$

and approximating all the a_{i1}'s by a referenced mean sound slowness (this approximtion has minimal effects on the model equations because all the resolved ray paths are near axial ray paths with small launching angles), the system for estimation can be expressed as

$$\underline{\delta t} = \underline{F}\ \underline{p} + \underline{v}, \tag{7.7a}$$

with

$$\underline{\delta t} = \begin{bmatrix} \underline{\delta t}_1 \\ \underline{\delta t}_2 \\ \vdots \\ \underline{\delta t}_N \end{bmatrix}, \quad \underline{F} = \begin{bmatrix} \underline{A}_1 & \underline{B} \\ \underline{A}_2 & \underline{B} \\ \vdots & \vdots \\ \underline{A}_N & \underline{B} \end{bmatrix}, \quad \underline{p} = \begin{bmatrix} \underline{\delta c} \\ \underline{\delta x} \end{bmatrix}, \quad \underline{v} = \begin{bmatrix} \underline{v}_1 \\ \underline{v}_2 \\ \vdots \\ \underline{v}_N \end{bmatrix}, \tag{7.7b}$$

where $\underline{A}_i$ is an qxn matrix with $\underline{w}_{i1}^T$ on its 1th row, and $\underline{B}$ is an qxu matrix with $\underline{b}_{i1}^T$ on its 1th row.

7.3 The Upper Erorr Variance Bound

We summarize the a priori information as follows: The parameters $\underline{\delta c}$ and $\underline{\delta x}$ to be estimated have zero means and a covariance matrix

$$\underline{C}_{\underline{p}} = \begin{bmatrix} \underline{C}_{\underline{\delta c}} & 0 \\ 0 & \sigma_x^2 \underline{I}_u \end{bmatrix} , \tag{7.8}$$

where $\underline{C}_{\underline{\delta c}}$ and $\sigma_x^2 \underline{I}_u$ are the covariance matrices of $\underline{\delta c}$ and $\underline{\delta x}$, respectively, and $\underline{\delta x}$ and $\underline{\delta c}$ are uncorrelated; $\underline{I}_u$ denotes an identity matrix with uxu dimension. For simplicity, all the δx_i's are assumed to be uncorrelated with each other and have the same variance σ_x^2. We further assume uncorrelated experimental noise with variance δ_v^2 such that the noise covariance matrix is

$$\underline{C}_{\underline{v}} = \sigma_v^2 \underline{I}_m . \tag{7.9}$$

We are now in the position to apply the generalized estimation procedure derived in Chapter 4, which is the minimization of the objective function $s(\underline{p})$ of (4.7), to the present situation. Since the model equations (7.7) are linear, the unique minimum of $s(\underline{p})$ at $\underline{p}=\underline{p}^*$ or $(\underline{\delta c},\underline{\delta x})=(\underline{\delta c}^*,\underline{\delta x}^*)$ is the linear minimum variance estimate, and its error covariance matrix is identical to the inverse Hessian matrix $\underline{H}^{-1}$ of $s(\underline{p})$. After replacing $\underline{F}$, $\underline{C}_{\underline{p}}$ and $\underline{C}_{\underline{v}}$ in the equation for $\underline{H}$ (4.22b) with their present definitions as given in (7.7b), (7.8) and (7.9), we obtain

$$\underline{H}^{-1} = \begin{bmatrix} \underline{C}_{\Delta\underline{\delta c}^*} & \underline{C}_{\Delta\underline{\delta c}^*,\Delta\underline{\delta x}^*} \\ \underline{C}^T_{\Delta\underline{\delta c}^*,\Delta\underline{\delta x}^*} & \underline{C}_{\Delta\underline{\delta x}^*} \end{bmatrix} = \begin{bmatrix} \underline{L}^{-1} & \sigma_v^{-2}(\sum_{i=1}^{N}\underline{A}_i^T)\underline{B} \\ \sigma_v^{-2}\underline{B}^T\sum_{i=1}^{N}\underline{A}_i & \sigma_x^{-2}\underline{I}_u + N\sigma_v^{-2}\underline{B}^T\underline{B} \end{bmatrix}^{-1} \tag{7.10}$$

with

$$\underline{L} = (\underline{C}^{-1}_{\underline{\delta c}} + \sigma_v^{-2}\sum_{i=1}^{N}\underline{A}_i^T\underline{A}_i)^{-1}, \tag{7.11}$$

where $\underline{C}_{\Delta\underline{\delta c}^*}$ and $\underline{C}_{\Delta\underline{\delta x}^*}$ are the error covariance matrices of $\underline{\delta c}^*$ and $\underline{\delta x}^*$, respectively, and $\underline{C}_{\Delta\underline{\delta c}^*,\Delta\underline{\delta x}^*}$ is the cross covariance matrix of the errors of $\underline{\delta c}^*$ and $\underline{\delta x}^*$. With the use of matrix identities given in standard mathematical texts for the inversions of block matrices, we further obtain

$$\underline{C}_{\Delta\underline{\delta c}^*} = [\underline{L}^{-1} - \frac{\sigma_v^{-2}}{N}(\sum_{i=1}^{N}\underline{A}_i^T)\underline{G}(\sum_{i=1}^{N}\underline{A}_i)]^{-1}, \tag{7.12}$$

where

$$\underline{G} = (\underline{B}\sigma_x)(\frac{\sigma_v^2}{N}\underline{I}_u + \sigma_x^2\underline{B}^T\underline{B})^{-1}(\underline{B}\sigma_x)^T. \tag{7.13}$$

Furthermore, from the equation for $\underline{p}^*$ (4.22a) with $\underline{p}_0=0$, $\underline{\delta c}^*$ can be equated to

$$\underline{\delta c}^* = \left(\frac{\underline{C}_{\Delta\delta c^*}}{\sigma_v^2}\right)\left[\sum_{i=1}^{N} \underline{A}_i^T \underline{\delta t}_i - \frac{1}{N}\left(\sum_{i=1}^{N} \underline{A}_i^T\right)\underline{G}\left(\sum_{i=1}^{N} \underline{\delta t}_i\right)\right]. \qquad (7.14)$$

Since $\underline{\delta x}^*$ and $\underline{C}_{\Delta\underline{\delta x}^*}$ are not our primary concerns here, their mathematical expressions are not presented.

It is seen from (7.11), (7.12) and (7.13) that for a given amount of a priori information ($\underline{C}_{\underline{\delta c}}$), a given noise level ($\sigma_v$), and a given tomographic configuration (which determines $\underline{A}_i$'s), $\underline{L}$ is the smallest error covariance matrix of $\underline{\delta c}^*$ that can be attained using known mooring motions. If the mooring motions are known so that σ_x=0 and hence $\underline{G}$=0, then $\underline{C}_{\Delta\underline{\delta c}^*}$=$\underline{L}$, and $\underline{C}_{\Delta\underline{\delta c}^*}$ increases as σ_x increases. However, in the limit when $\underline{\delta x}$ is large enough so that the ratios of σ_v^2 to the variances of the signals produced by $\underline{\delta x}$ (the diagonal elements of $\sigma_x^2\underline{B}^T\underline{B}$) approach zero, $C_{\Delta\underline{\delta c}^*}$ approaches its maximum bound $\underline{U}$ and it becomes invariant with σ_x because $\underline{G}$ approaches

$$\underline{G}_U = \underline{B}\,\underline{B}^+, \qquad (7.15)$$

where $\underline{B}^+$ is the pseudoinverse of $\underline{B}$, and $\underline{G}$ is no longer a function of σ_x. This upper error variance bound $\underline{U}$ of $\underline{\delta c}^*$ can be expressed as

$$\underline{U} = \left[\underline{L}^{-1} - N^{-1}\sigma_v^{-2}\left(\sum_{i=1}^{N} \underline{A}_i^T\right)\underline{G}_U\left(\sum_{i=1}^{N} \underline{A}_i\right)\right]^{-1}. \qquad (7.16)$$

In this limit, $\underline{\delta c}^*$ is also independent of σ_x and can be expressed as

$$\underline{\delta c}_U^* = \left(\frac{\underline{U}}{\sigma_v^2}\right)\left[\sum_{i=1}^{N} \underline{A}_i^T \underline{\delta t}_i - \frac{1}{N}\left(\sum_{i=1}^{N} \underline{A}_i^T\right)\underline{G}_U\left(\sum_{i=1}^{N} \underline{\delta t}_i\right)\right]. \qquad (7.17)$$

7.4 The Differenced System

The differenced system can be expressed as

$$\begin{bmatrix} \underline{\delta t}_2-\underline{\delta t}_1 \\ \underline{\delta t}_3-\underline{\delta t}_1 \\ \vdots \\ \underline{\delta t}_N-\underline{\delta t}_1 \end{bmatrix} = \begin{bmatrix} \underline{A}_2-\underline{A}_1 \\ \underline{A}_3-\underline{A}_1 \\ \vdots \\ \underline{A}_N-\underline{A}_1 \end{bmatrix} \underline{\delta c} + \begin{bmatrix} \underline{v}_2-\underline{v}_1 \\ \underline{v}_3-\underline{v}_1 \\ \vdots \\ \underline{v}_N-\underline{v}_1 \end{bmatrix} \tag{7.18}$$

in which $\underline{\delta x}$ is eliminated. Notice that the elimination is done by subtracting a set of model equations ($\underline{\delta t}_1=\underline{A}_1\underline{\delta c}+\underline{v}_1$) from the other sets ($\underline{\delta t}_i=\underline{A}_i\underline{\delta c}+\underline{v}_i$). The corresponding estimation is therefore based on the differenced data ($\underline{\delta t}_i-\underline{\delta t}_1$), the differenced model equations ($\underline{A}_i-\underline{A}_1$) and the differenced noises ($\underline{v}_i-\underline{v}_1$). The noise of the new system (7.18) is correlated and has twice the variance of the original system (7.7). The covariance matrix of the differenced noises is

$$\underline{C}_{\Delta v} = \sigma_v^2 \left[\underline{I}_m + \begin{pmatrix} \underline{I}_q & \cdots & \underline{I}_q & \cdots & \underline{I}_q \\ \vdots & & \vdots & & \vdots \\ \underline{I}_q & \cdots & \underline{I}_q & \cdots & \underline{I}_q \end{pmatrix} \right]. \tag{7.19}$$

Applying (4.22a) and (4.22b) to the differenced system, and equating, the error covariance matrix $\underline{U}_\Delta$ of the estimated $\underline{\delta c}$ and the estimate $\underline{\delta c}_\Delta{}^*$ itself become

$$\underline{U}_\Delta = [\underline{L}^{-1} - \frac{\sigma_v^{-2}}{N} (\sum_{i=1}^{N} \underline{A}_i^T)(\sum_{i=1}^{N} \underline{A}_i^T)]^{-1} \tag{7.20}$$

and

$$\underline{\delta c}^*_{\Delta} = \left(\frac{\underline{U}_{\Delta}}{\sigma_v^2}\right) \left[\sum_{i=1}^{N} \underline{A}_i^T \underline{\delta t}_i - \frac{1}{N} \left(\sum_{i=1}^{N} \underline{A}_i^T \right) \left(\sum_{i=1}^{N} \underline{\delta t}_i \right) \right]. \qquad (7.21)$$

Interestingly, if the product q of the number of sources and receivers coincides with the rank of B when $q \leq u$, that is when B is underdetermined, then we have $G_U = I$ and hence $U = U_{\Delta}$ and $\underline{\delta c}^* = \underline{\delta c}_{\Delta}^*$. It is always true that $\underline{U} \leq \underline{U}_{\Delta}$, and in fact, if a lot of moorings are deployed so that $q >> u$ and hence the diagonal elements of $\underline{G}_U$ are significantly smaller that unity, then $\underline{U} << \underline{U}_{\Delta}$. However, in realistic experiments, $q \sim u$, implying that $\underline{U} \sim \underline{U}_{\Delta}$.

It is found that the error variance of $\underline{\delta c}^*$ for given noise level, a priori information and geometry of the acoustic array is bounded approximately between $\underline{L}$ and $\underline{U}_{\Delta}$, as given in (7.11) and (7.20), respectively, and the error variance approaches the upper bound $\underline{U}_{\Delta}$ as σ_x increases. If in practice σ_x always exceeded a critical value such that $\underline{U}_{\Delta}$ is always reached, then $\underline{U}_{\Delta}$ can be used as a guideline in the determination of whether the tracking of mooring motion is needed for a given experimental setup. The crucial question, therefore, is to find how small that critical value of σ_x is or how large can σ_x be before the upper bound is reached. We will pursue the answer through a computer-simulation study next.

7.5 Numerical Results

Computer simulations are used here to study how large σ_x^2 can be before $\underline{C}_{\underline{\Delta\delta c}^*}$ reaches $\underline{U} \sim \underline{U}_\Delta$ for a typical situation. The tomographic configuration and 58 ray paths of the 1981 experiment are used in this simulated study; there are q=NSxNR=4x5=20 S-R pairs, u=2(NS+NR)=2(4+5)=18 unknown δx_i's and about three ray paths used per S-R pair; the rank of $\underline{B}$ is 18-3=15. The vertical structure of the simulated δc consists of only the first baroclinic perturbation. Horizontally, the simulated δc is homogenous and isotropic, and has a Gaussian correlation function with a decay scale of 100 km and an rms value of 2 m/s at a depth of 700 m. The noise variance σ_v^2 is set to 5^2 ms^2.

The covariance matrices $\underline{C}_{\underline{\Delta\delta c}^*}$ for σ_x=0, 100 m and 200 m, and the upper error variance bound $\underline{U}_\Delta$ are calculated numerically. The standard deviation of δc^* (i.e. square root of the diagonal elements of $\underline{C}_{\underline{\Delta\delta c}^*}$) for σ_x=0 and 200 m, and the upper bound for the standard deviation (i.e. square root of the diagonal elements of $\underline{U}_\Delta$) at a depth of 700 m are contoured in Figures (7.1), (7.2) and (7.3), respectively. The rms errors of $\underline{\delta c}^*$ versus σ_x at two representative locations (a) and (b) in space are also plotted in Figures (7.4a) and (7.4b), respectively. While (a) is located in an area with a low density of ray paths at the lower right corner of the experimental region, (b) is located in an area with a high

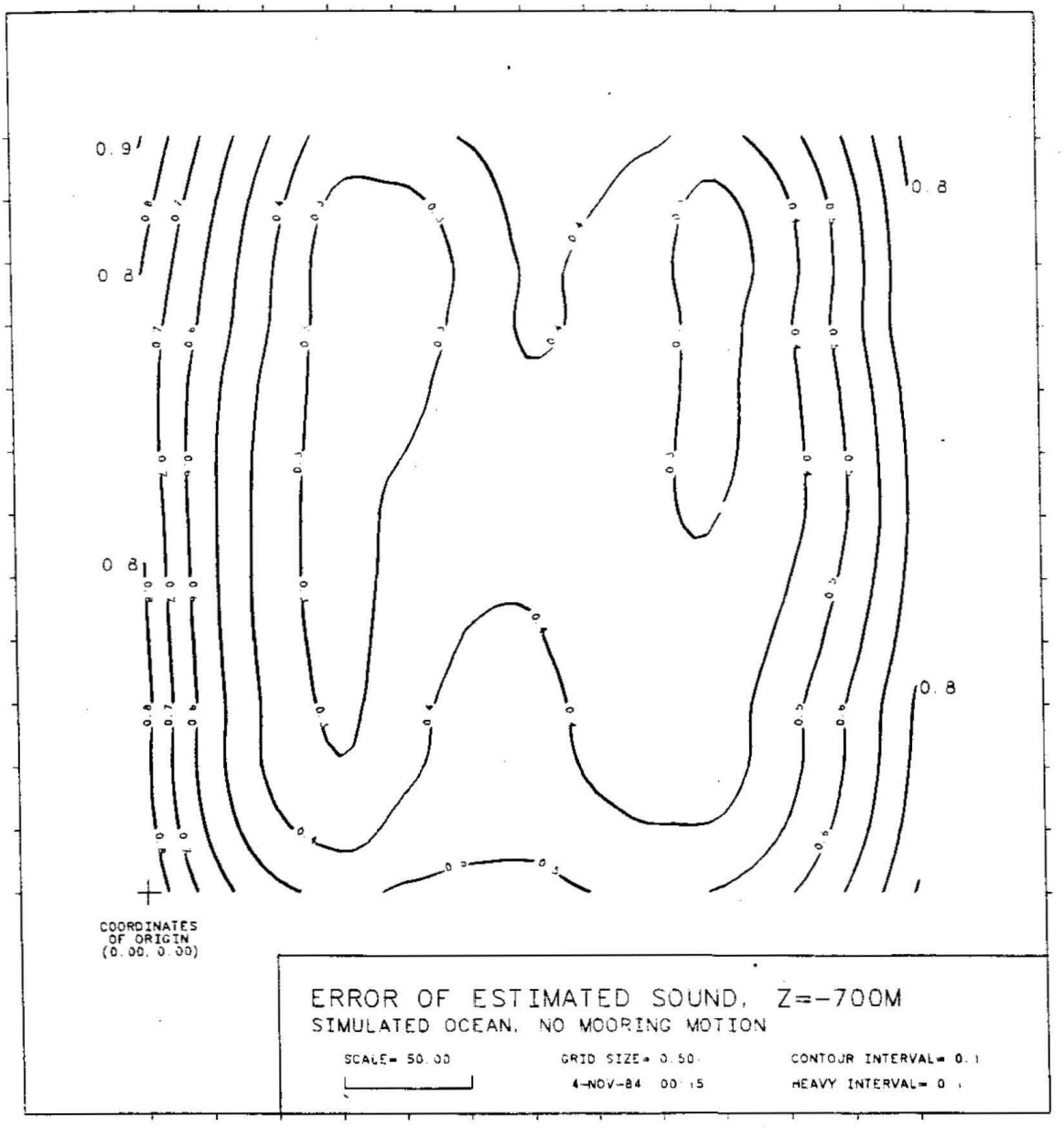

Figure 7.1. Standard deviation, at a depth of 700 m in the tomographic square, of the linear, tomographic sound-speed perturbation estimate in the absence of untracked mooring motion. The sound-speed perturbation has an rms value of 2 m/s and a horizontal correlation length of 100 km. The experimental noise is 5 ms (rms). Contour interval is 0.1 m/s.

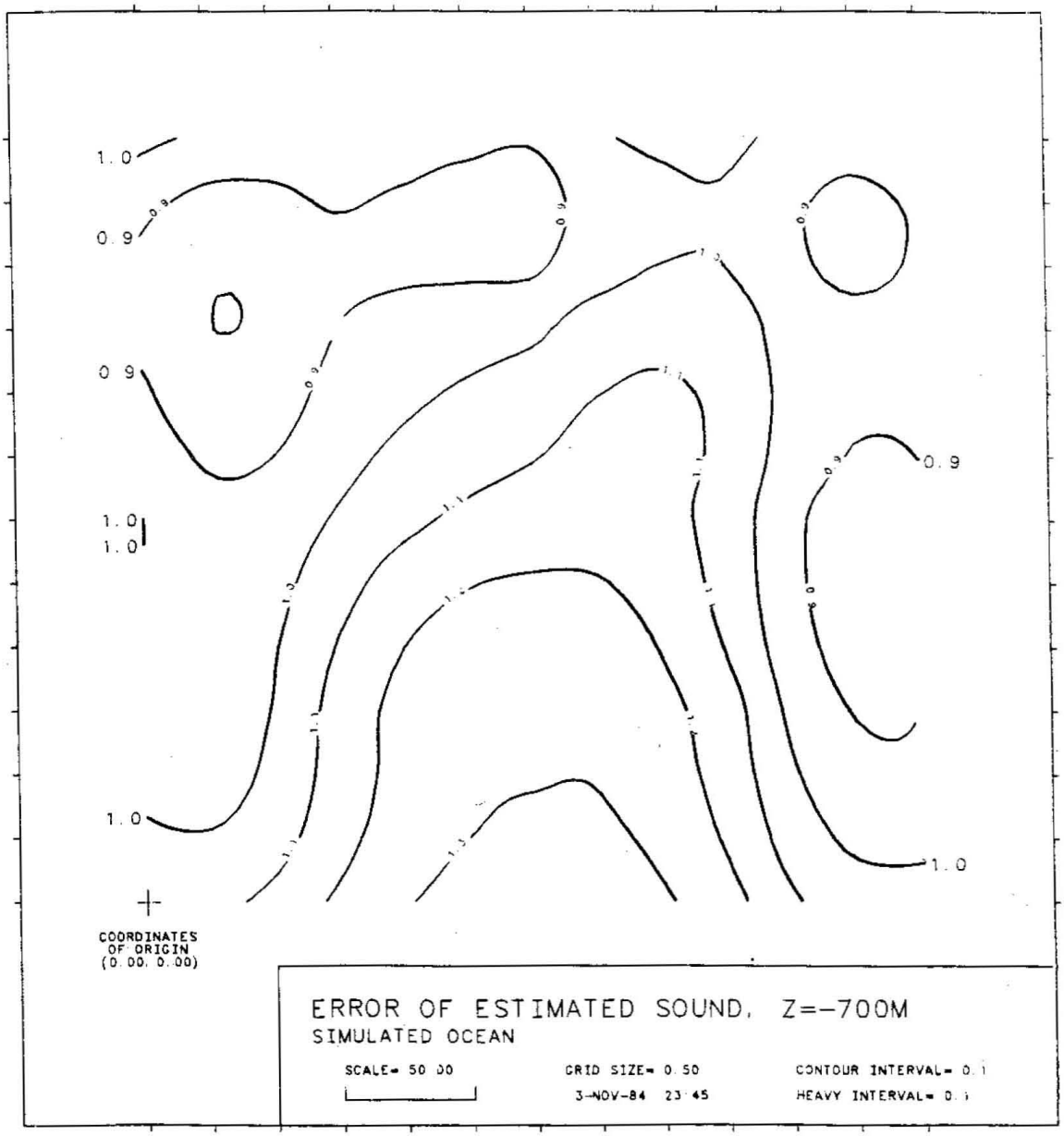

Figure 7.2. Standard deviation, at a depth of 700 m in the tomographic square, of the linear, tomographic sound-speed perturbation estimate in the presence of 200 m (rms) untracked mooring motion. The sound-speed perturbation has an rms value of 2 m/s and a horizontal correlation length of 100 km. The experimental noise is 5 ms (rms). Contour interval is 0.1 m/s.

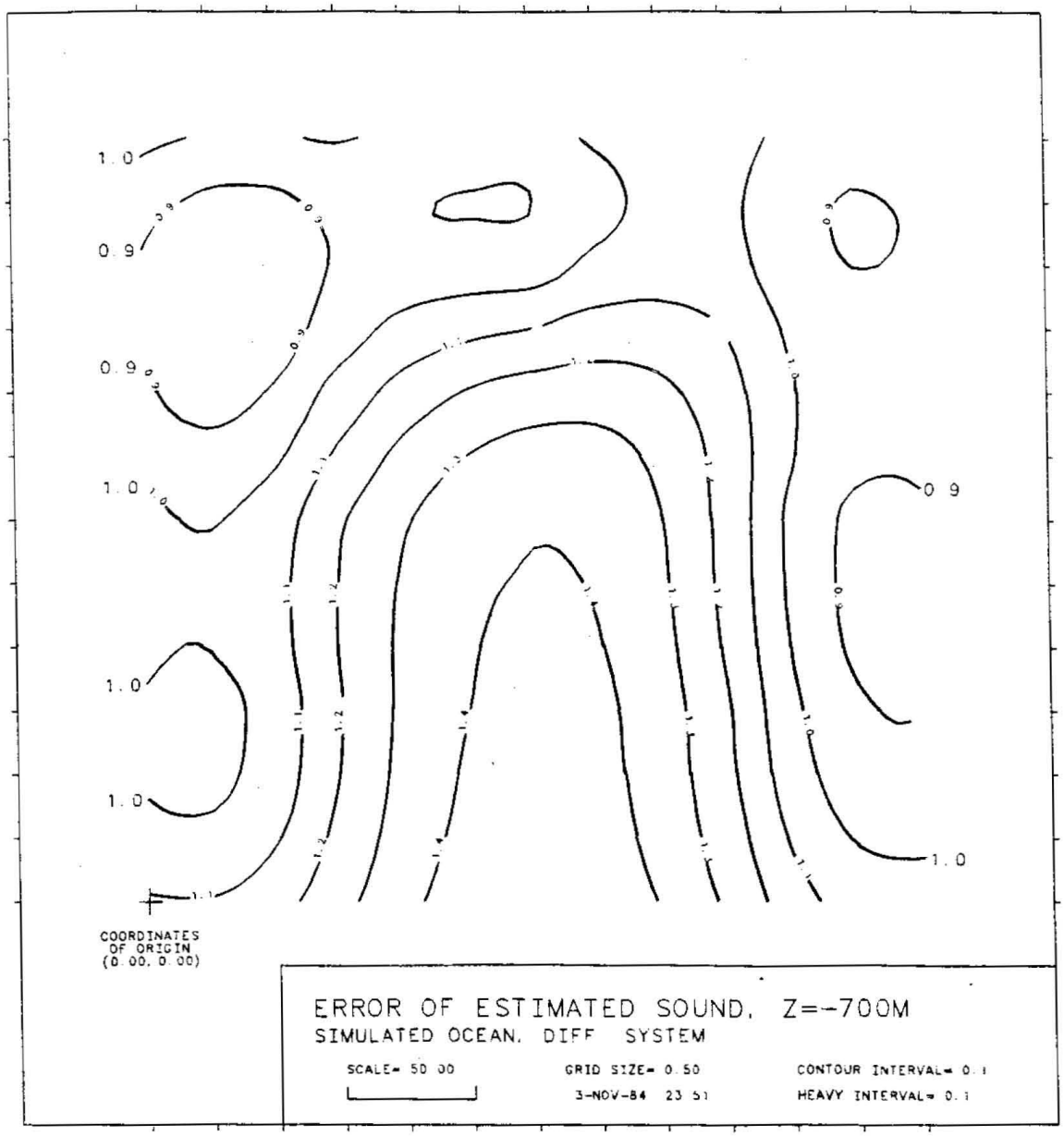

Figure 7.3. Standard deviation, at a depth of 700 m in the tomographic square, of a linear, tomographic sound-speed perturbation estimate obtained from the differenced system. These errors are approximately the upper-bound errors. The sound-speed perturbation has an rms value of 2 m/s and a horizontal correlation length of 100 km. The experimental noise is 5 ms (rms). Contour interval is 0.1 m/s.

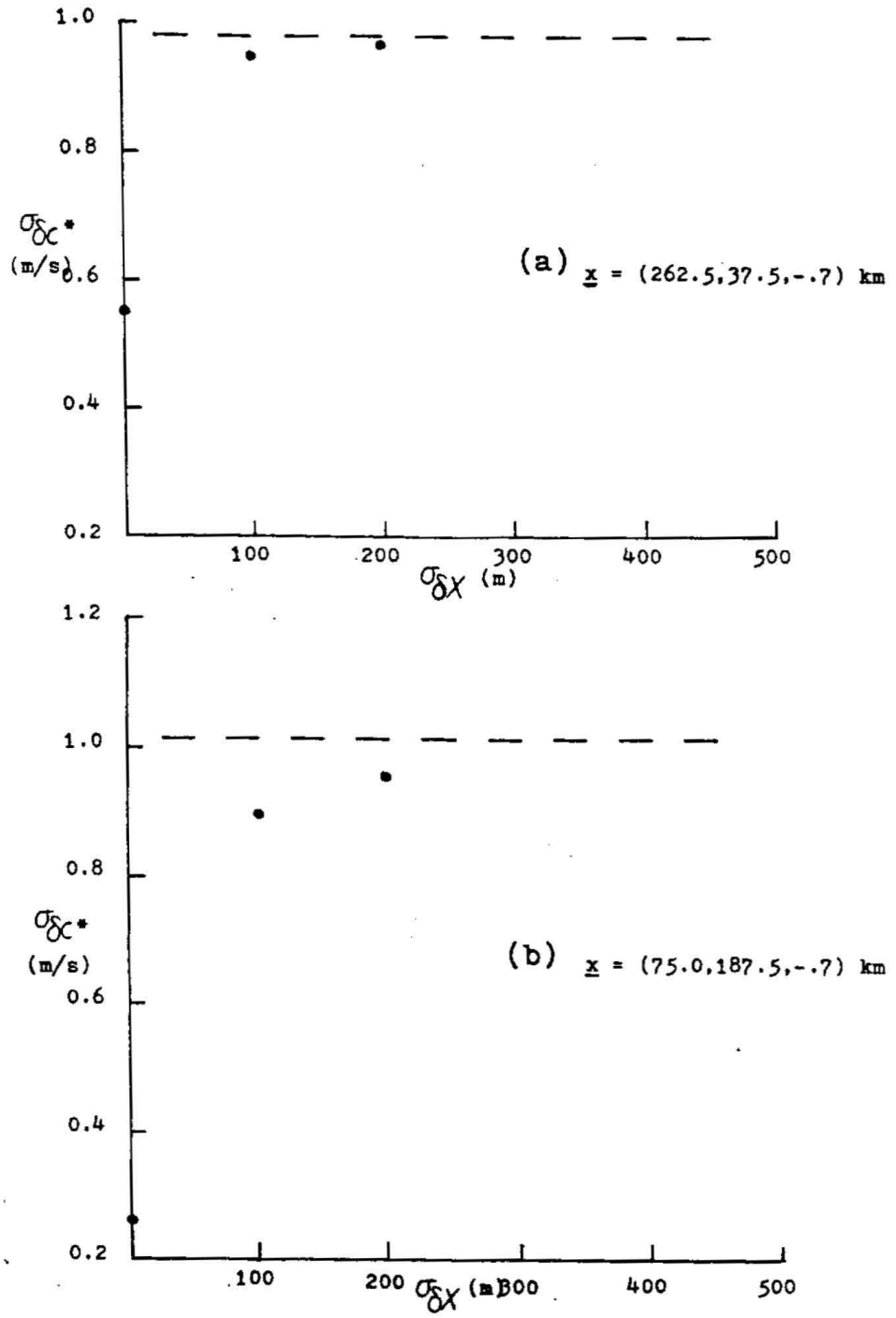

Figure 7.4a and b. The dependence of the standard deviation of the linear, tomographic sound-speed perturbation estimate at two locations on rms mooring displacement in the absence of mooring tracking. The figures show that the upper standard-deviation bound (_____) is rapidly reached. The upper bound shown is approximated from the differenced system. The sound-speed perturbation has an rms value of 2 m/s and a horizontal correlation length of 100 km. The experimental noise is 5 ms (rms).

density of ray paths near the center of the region. The upper bounds for the standard deviation of δc^* at the two locations are also plotted in the corresponding figures.

It is seen from the figures that the error converges very rapidly to the upper bound; the standard deviation of $\underline{\delta c}^*$ for a small σ_x of only 200 m is nearly equal to the maximum standard deviation. For this particular experiment, it is indicated from Figures (7.1) and (7.3) that in order to estimate $\underline{\delta c}$ accurately, say to within ±0.5 m/s, tracking mooring motions is required. Notice that the regions with more ray paths passing through them have smaller errors only when $\sigma_x=0$, that is only when the oceanic signals are dominant in the data. This is because as far as the estimation of δc is concerned, noise becomes dominant in the data when mooring motions are not tracked, and since the regions with higher density of ray paths resolve more data variance, they also resolve more noise variance in this case.

7.6 Conclusions

The error variance of the estimated δc very rapidly reaches an upper bound as σ_x^2 increases. When the differenced system is used, the upper error variance bound and the associated estimate of δc coincide (or approximately coincide) with the error variance and the estimate in the estimation process (unless $q>>u$). Therefore, the decision of whether to deploy navigational systems for tracking mooring motions in a particular experiment can be made simply by a simulated study of the error variance associated with the differenced system alone, and if this rms error is not tolerable then tracking mooring motions must be used. The upper error bound can be lowered by reducing the noise level or increasing the number of sourses and receivers, and these are the alternatives to tracking mooring motions when a good estimate of δc is desired.

APPENDIX

A DERIVATION OF THE SEQUENTIAL-REESTIMATION ALGORITHM

Let us choose the free parameters of the system (6.11) to be the spectral amplitudes $\underline{a}(t_N)$ at the final time t_N in the sequence of observations and define the functions $s^{(1)}$ of $\underline{a}(t_1)$ by

$$s^{(1)}[\underline{a}(t_1)] \equiv \sum_{k=0}^{1} s_k[\underline{a}(t_k)], \tag{A.1a}$$

where

$$s_0[\underline{a}(t_0)] = 1/2\ [\underline{\delta t}^{\rho}(t_0)-\underline{A}\ \underline{a}(t_0)]^T \underline{C}_{\underline{v}}^{-1}(t_0)\ [\underline{\delta t}^{\rho}(t_0)-\underline{A}\ \underline{a}(t_0)] + 1/2\ \underline{a}(t_0)^T \underline{C}_{\underline{a}}^{-1}(t_0)\ \underline{a}(t_0), \tag{A.1b}$$

$$s_k[\underline{a}(t_k)] = 1/2\ [\underline{\delta t}^{\rho}(t_k)-\underline{A}\ \underline{a}(t_k)]^T \underline{C}_{\underline{v}}^{-1}(t_k)\ [\underline{\delta t}^{\rho}(t_k)-\underline{A}\ \underline{a}(t_k)] \quad \text{for } k>0, \tag{A.1c}$$

and $\underline{a}(t_k)$ with $k<1$ is linearly related to $a(t_1)$ according to the linear dynamical relation (6.13). In (A.1b), $\underline{C}_{\underline{a}}(t_0)$ represents the a priori covariance of $\underline{a}(t)$ at t_0 or any other time. With the noise being uncorrelated at different times, it follows from the objective-function approach that the minimum-variance, maximum-likelihood estimate $\underline{a}^*(t_N)$ of $a(t_N)$ can be evaluated by minimizing

$$s[\underline{a}(t_N)] = s^{(N)}[\underline{a}(t_N)]. \tag{A.2}$$

Through a Taylor-series expansion, we can recast the quadratic function (A.1) as

$$s^{(l)}[\underline{a}(t_l)] = s^{(l)}[\underline{a}^*(t_l)] + 1/2\ [\underline{a}(t_l)-\underline{a}^*(t_l)]^T \underline{H}_l [\underline{a}(t_l)-\underline{a}^*(t_l)] \tag{A.3}$$

where $\underline{a}^*(t_l)$ is the minimum point and $\underline{H}_l$ is the Hessian matrix of $s^{(l)}$. Furthermore, through the use of (6.13), (A.1a) and (A.3), $s^{(l+1)}$ can be expressed as

$$s^{(l+1)}[\underline{a}(t_{l+1})] = \frac{1}{2}[\underline{a}(t_{l+1})-\underline{a}^p(t_{l+1})]^T \underline{C}_{\underline{a}}^{-1}(t_{l+1})\ [\underline{a}(t_{l+1})-\underline{a}^p(t_{l+1})] + s_{l+1}[\underline{a}(t_{l+1})], \tag{A.4a}$$

where

$$\underline{a}^p(t_{l+1}) = \underline{D}_l \underline{a}^*(t_l) \tag{A.4b}$$

and

$$\underline{C}_{\underline{a}}(t_{l+1}) = \underline{D}_l \underline{H}_l^{-1} \underline{D}_l^T. \tag{A.4c}$$

We have neglected the constant term $s^{(l)}[a^*(t_l)]$ in writing down (A.4a); this is of no consequence in the subsequent minimization of

$s^{(l+1)}$. After setting the gradient of $s^{(l+1)}$ to zero, the unique minimum of $s^{(l+1)}$ is found to be at

$$\underline{a}^*(t_{l+1}) = \underline{H}_{l+1}^{-1}[\underline{A}^T \underline{C}_v^{-1}(t_{l+1})\underline{\delta t}^o(t_{l+1}) + \underline{C}_a^{-1}(t_{l+1})\underline{a}^p(t_{l+1})], \tag{A.5a}$$

where

$$\underline{H}_{l+1}^{-1} = \underline{C}_a(t_{l+1}) - [\underline{C}_a(t_{l+1})\underline{A}^T][\underline{A}\ \underline{C}_a(t_{l+1})\underline{A}^T + \underline{C}_v(t_{l+1})]^{-1}[\underline{C}_a(t_{l+1})\underline{A}^T]^T. \tag{A.5b}$$

It is now clear that the optimal estimate $\underline{a}^*(t_N)$ can be obtained by computing the $\underline{a}^*(t_l)$'s, that is sequentially minimizing the functions $s^{(l)}$ in order of increasing l. Each minimization in the sequence can be interpreted as an improved reestimation of the field. The covariance of the field is updated at each time step of the reestimation process by the information gained from the preceeding minimizations. At the $(l+1)$th time step, using (A.4b), (A.4c) and (A.5), a prediction $\underline{a}^p(t_{l+1})$ of $\underline{a}(t_{l+1})$ is first extrapolated from $\underline{a}^*(t_l)$ which, on the other hand, is an estimate of $a(t_l)$ based on the data obtained prior to t_{l+1}; then the predicted value is corrected in an estimation that uses the updated covariance $\underline{C}_a(t_{l+1})$ and the data obtained at t_{l+1}.

REFERENCES

Acton, F.S., "NUMERICAL METHODS THAT WORK," Happer and Row, New York, 1970.

Backus, G.E. and J.F. Gilbert, "Numerical applications of a formalism for geophysical inverse problem," Geophys. J. Roy. astr. Soc., 13, 247-276, 1967.

Backus, G.E. and J.F. Gilbert, "The resolving power of gross Earth data," Geophys. J. R. astr. Soc., 16, 169-205, 1968.

Backus, G.E. and J.F. Gilbert, "Uniqueness in the inversion of inaccurate gross Earth data," Phil. Trans. R. Soc. London Ser. A, 266, 123-192, 1970.

Bard, Y., "Nonlinear Parameter Estimation," New York: Academic, 1974.

Bendat, J.S. and A.G. Piersol, "RANDOM DATA: ANALYSIS AND MEASUREMENT PROCEDURES," John Wiley and Sons, Inc., 1971.

Blokhintev, I., "Acoustics of a Nonhomogenous Moving Medium," Nat. Adv. Comm. Aeronaut. Tech. Mem., 1399, 1965.

Box, G.E.P. and H.L. Lucas, "Design of experiments in nonlinear situations," Biometrika, 46, 77-90, 1959.

Box, G.E.P. and W.G. Hunter, "Sequential design of experiments for nonlinear models," Proc. IBM Sci. Comput. Symp. Statist., IBM, White Plains, New York, 1963.

Box, G.E.P. and W.J. Hill, "Discrimination among mechanistic models," Technometrics, 9, 57-71, 1967.

Brekhovskikh L.M., K.N. Fedorov, L.M. Fomin, M.N. Koshlyakov and A.D. Yampolsky, "Large-scale multi-buoy experiment in the tropical Atlantic," Deep-Sea Res., 18, 1189-1206, 1971.

Bretherton, F., R.E. Davis and C.B. Fandry, "A technique for objective analysis and design of experiments applied to MODE-73," Deep-Sea Res., 23, 559-582, 1976.

Clay, C.S. and H. Medwin, "Acoustical Oceanography: Principle and Applications," John Wiley and Sons, 1977.

Cornuelle B.D., "Acoustic Tomography," IEEE Transactions on Geoscience and Remote Sensing, Volume GE-20, 326-332, 1982.

Cornuelle, D.B., "IVERSE METHOD AND RESULTS FROM THE 1981 OCEAN ACOUSTIC TOMOGRAPHY EXPERIMENT", Woods Hole Oceanographic Inst./Mass. Inst. Of Tech., Ph. D. thesis, 1983.

Cornuelle B., C. Wunsch, D. Behringer, T. Birdsall, M. Brown, R. Heinmiller, R. Knox, K. Metzger, W. Munk, J. Spiesberger, R. Spindel, D. Webb, P. Worcester, "Tomographic Maps of the Ocean Mesoscales - 1: Pure Acoustics," J. Phys. Oceanography, in press, 1985.

Dahlquist G. and A. Bjorck, "Numerical Methods," (translated by N. Anderson,) Prentice-Hall Inc., 1974.

Drake, A.W., "fundamentals of applied probability theory," McGraw-Hill, 1967.

Fisher, R.H., "Contributions to Mathematical Statistics," (collection of papers published in 1920-1943,) Wiley, New York, 1950.

Flatte, S.M. (editor), "SOUND TRANSMISSION THROUGH A FLUCTUATING OCEAN," Cambridge University Press, 1979.

Fletcher, R. and M.J.D. Powell, "A rapidly convergent descent method for minimization",The Computer Journal, Volume 6, 1983.

Flierl, G.R., "MODELS OF VERTICAL STRUCTURE AND CALIBRATION OF TWO-LAYER MODELS," Dynamics of Atmospheres and Oceans, Volume 2, 341-381, 1978.

Gelb, A. (editor), "Applied Optimal Estimation," M.I.T. Press, 1982.

Hamilton, G., W.L. Siegmann and M.J. Jacobson, "Simplified calculation of ray-phase perturbations due to ocean-enviromental variations," J. Acoust. Soc. Am., 67, 1980.

Hogg, N.G., "Observations of Internal Kelvin Waves Trapped Round Bermuda," Am. Meteo. Soc., 10, 1980.

Jackson, D.D., "Interpretation of Inaccurate, insufficient and inconsistent data," J. Roy. Astron. Soc., 28, 97-100, 1972.

Jackson, D.D., "The use of a priori data to resolve non-uniqueness in linear inversion", Geophys. J. R. astr. Soc., 57, 1979.

Jenkins, G.M. and D.G. Watts, "SPECTRAL ANALYSIS and its applications," Holden-Day, 1968.

Koshlyakov M.N. and Y.M. Grachev, "Mesoscale currents at a hydrophysical polygon in the tropical Atlantic," Deep-Sea Res., 20, 507-526, 1973.

Lanczos, C., "Linear Differential Operators," Van Nostrand, New York, 1961.

LeBlond, P.H. and L.A. Mysak, "Waves In The Ocean," Elsevier Scientific Publishing Company, 1978.

Liebelt, P.B., "An introduction to Optimal Estimation," Addison Wesley, Reading Ma., 1967.

Longuet-Higgins M.S., F.R.S. and A.E. Gill, "Resonant interactions between planetary waves," Proc. R. Soc. Lond., A229, 773-784, 1967.

Longuet-Higgins M.S., "On the trapping of long-period waves round islands," J. Fluid Mech., vol. 37, part 4, 120-145, 1969.

Lovett, J.R., "Merged seawater sound-speed equations," J. Acoust. Soc. Am., 63, 1713, 1978.

Mercer, J.A. and J.R. Booker, "Long-range propagation of sound through oceanic mesoscale structures," J. Geophys. Res., 88, 689-700, 1983.

Mcwilliams J. and A. Robinson, "A wave analysis of the Polygon array in the tropical Atlantic," Deep-Sea Res., 21, 359-368, 1974.

McWilliams, J.C. and G.R. Flierl, "Quasigeostrophic wave analyses," (in "Dynamics and the analysis of MODE-1," editor: A.R. Robinson,) M.I.T., Cambridge, 54-1417, 1975.

McWilliams, J.C. and G.R. Flierl, "Optimal, quasi-geostrophic wave analyses of MODE array data," Deep-Sea Research, 23, 285-300, 1976.

Medwin, H., "Speed of sound in water: a simple equation for realistic parameters," J. Acoust. Soc. Am., 58, 1318-1319, 1978.

MODE Group, "The Mid-ocean dynamics experiment," Deep-Sea Res., 25, 859-910, 1978.

Mooers, C.N.K., "Sound-velocity perturbations due to low-frequency motions in the ocean," J. Acoust. Soc. Am., 57, 1067-1075, 1975.

Munk, W. and C. Wunsch, "Ocean acoustic tomography: a scheme for largescale monitoring," Deep-Sea Research, 26A, 123-161, 1979.

Munk, W., "Horizontal deflection of acoustic paths by mesoscale eddies," J. Phys. Oceanography, 10, 596-604, 1980.

Munk, W. and C. Wunsch, "Observing the ocean in 1990s," Phil. Trans. R. Soc. Lond., A307, 439-464, 1982.

New, R., T. Eisler, D. Calderone and D. Porter, "OCEAN ACOUSTIC TOMOGRAPHY: A PRELIMINARY EVALUATION," System Analysis Staff, Office of Ocean Technology and Engineering Services, and National Oceanic and Atmospheric Administration, report, 1982.

Parker, R.L., "UNDERSTANDING INVERSE THEORY," Ann. Rev. Earth Planetory Science, Volume 5, 1977.

Pedlosky, J., "Geophysical Fluid Dynamics," New York, Springer-Verlag, 1980.

Pickard, G.L. and W.J. Emery, "Descriptive Physical Oceanography: An Introduction," Fourth Enlarged Edition, Pergamon Press, 1982.

Piips, T., "Precision Sound Velocity Profiles in the Ocean, vol. II-The Sound Channel in the Bermuda-Barbados Region, April-July 1964", Tech. Rep. No. 4, Lamont Geological Observatory, Columbia U., Palisades, NY, 1967.

Provost, C., "A Variational Method for Estimating the General Circulation in the Ocean," U. of Cal. at San Diego, Ph. D. thesis, 1983.

Rhines, P.B., "Waves and turbulence on a beta-plane," J. Fluid Mech., 69, 417-433, 1975.

Richman, J.G., C. Wunsch and N.G. Hogg, "Space and Time Scales of Mesoscale Motion in theWestern North Atlantic", Rev. Geophysics and Space Physics,Volume 15, 385-420, 1977.

Shannon, C.E., "A mathematical theory of communication," Bell System Tech. Journal, 27, 623-656, 1948.

Rust, B.W. and W.R. Burrus, "Mathematical programming and the numerical solution of linear equations," New York, American Elsevier Pub. Co., 1972.

Sanford, T.B., "Observations of the Vertical Structure of Internal Waves," J. Geophy. Res., Vol. 80, no. 27, 3861-3871, 1975.

Steiglitz, K., "AN INTRODUCTION TO DISCRETE SYSTEMS," John Wiley and Sons, Inc., 1974.

Spiesberger, J.L., R.C. Spindel and K. Metzger, "Stability and identification of long range ocean acoustic multipaths," J. Acoust. Soc. Am., 67, 2011-2017, 1980.

Spindel, R.C. and Y. Desaubies, "Eddies and Acoustics," (in "Eddies in Marine Science," editor: A.R. Robinson) Springer-Verlag Berlin Heidelberg, 1983.

Spindel, R.C., "An Underwater Acoustic Pulse Compression System," IEEE Tran. Acoustic, Speech and Signal Proc., Vol. ASSP-27, NO.6, 723-728, 1979.

Spindel R.C. and J.L. Spiesberger, "Multipath variability due to the Gulf Stream," J. Acoust. Soc. Am., 69, 982-988, 1981.

The Ocean Tomography Group, "A demonstration of ocean acoustic tomography," Nature, 299, 121-125, 1982.

Ugincius, P., "Ray Acoustics and Fermat's Principle in a Moving Inhomogenous Medium," J. Acoust. Soc. Am., 51, 1759-1763, 1970.

U.S. POLYMODE Organizing Committee, "U.S. POLYMODE- Program and Plan," Dept. Meteorology, M.I.T., 1976.

Wilson, W., "Equation for the Speed of Sound in Seawater," J. Acoust. Soc. Am., 32, 1357, 1960.

Wiggins, R.A., "The General Linear Inverse Problem: Implication of Surface Waves and Free Oscillations for Earth Structure," Rev.Geophysics and Space Physics, Volume 10, 1972.

Wunsch, C., "The general circulation of the North Atlantic west of 50W determined from inverse methods," Rev. Geophysics and Space Physics, Volume 16, 1978.

Wunsch, C., "The spectrum from two years to two minutes of temperature fluctuations in the main thermocline at Bermuda," Deep-Sea Res., 19, 577-593, 1972.